How tissues tell time

Exploring transcriptional controls for phase- and tissue-specific circadian gene activities in peripheral cells

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Dr. rer. nat.
im Fach Biologie

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Dipl.-Chem. Agnes Lioba Rosahl

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Stefan Hecht, Ph.D.

Gutachter:
1. Prof. Dr. Hanspeter Herzel
2. Prof. Dr. Achim Kramer
3. Prof. Dr. Uwe Ohler

eingereicht am: 25.3.2014
Tag der mündlichen Prüfung: 29.9.2014

Für
meine Mutter
und
meine Tochter

Abstract

A circadian clock in peripheral tissues regulates physiological functions through gene expression timing. However, despite the common and well studied core clock mechanism, understanding of tissue-specific regulation of circadian genes is marginal.

Overrepresentation analysis is a tool to detect transcription factor binding sites that might play a role in the regulation of co-expressed genes. To apply it to circadian genes that do share a period of about 24 hours, but differ otherwise in peak phase timing and tissue-specificity of their oscillation, clear definition of co-expressed gene subgroups as well as the appropriate choice of background genes are important prerequisites. In this setting of multiple subgroup comparisons, a hierarchical method for false discovery control reveals significant findings.

Based on two microarray time series in mouse macrophages and liver cells, tissue-specific regulation of circadian genes in these cell types is investigated by promoter analysis. Binding sites for CLOCK:BMAL1, NF-Y and CREB transcription factors are among the common top candidates of overrepresented motifs. Related transcription factors of BHLH and BZIP families with specific complexation domains bind to motif variants with differing strengths, thereby arranging interactions with more tissue-specific regulators (e.g. HOX, GATA, FORKHEAD, REL, IRF, ETS regulators and nuclear receptors). Presumably, this influences the timing of pre-initiation complexes and hence tissue-specific transcription patterns.

In this respect, the content of guanine (G) and cytosine (C) bases as well as CpG dinucleotides are important promoter properties directing the interaction probability of regulators, because affinities with which transcription factors are attracted to promoters depend on these sequence characteristics.

Zusammenfassung

Die zirkadiane Uhr reguliert physiologische Funktionen in vielen Organen durch zeitlich gesteuerte Genexpression. Obwohl der zugrundeliegende allgemeine Uhrmechanismus bereits recht gut untersucht ist, bestehen noch viele Unklarheiten über die gewebespezifische Regulation zirkadianer Gene. Diese haben zwar eine Periode von etwa 24 Stunden im Expressionsmuster gemeinsam, unterscheiden sich aber ansonsten darin, zu welcher Tageszeit sie am höchsten exprimiert sind und in welchem Gewebe sie oszillieren.

Überrepräsentationsanalyse ist eine Methode, um Bindungsstellen von Transkriptionsfaktoren zu identifizieren, die in der Regulation ähnlich exprimierter Gene eine Rolle spielen könnten. Um sie auch auf zirkadiane Gene anzuwenden, ist es nötig, Untergruppen ähnlich exprimierter Gene genau zu definieren und andere Gene passend zum Vergleich auszuwählen. Eine hierarchische Methode zur Kontrolle des Anteils falscher Entdeckungen hilft, aus der daraus entstehenden Menge vieler Untergruppenvergleiche signifikante Ergebnisse zu filtern.

Basierend auf mit Microarrays gemessenen Zeitreihen aus Makrophagen und Leberzellen von Mäusen wurde durch Promotoranalyse die gewebespezifische Regulation von zirkadianen Genen in diesen beiden Zelltypen untersucht. Bindungsstellen der Transkriptionsfaktoren CLOCK:BMAL1, NF-Y und CREB fanden sich in beiden als überrepräsentiert. Verwandte Transkriptionsfaktoren der BHLH und BZIP Familien mit spezifischen Komplexierungsdomänen binden mit unterschiedlicher Stärke an Motivvarianten und arrangieren dabei Interaktionen mit gewebespezifischeren Regulatoren (z. B. HOX, GATA, FORKHEAD, REL, IRF, ETS Regulatoren und nukleare Rezeptoren). Es ist anzunehmen, daß das den Zeitablauf der Komplexbildung am Promotor zum Start der Transkription beeinflußt und daher auch die gewebespezifischen Transkriptionsmuster.

In dieser Hinsicht sind der Gehalt von Guanin (G) und Cytosin (C) sowie deren CpG-Dinukleotiden wichtige Promotoreigenschaften, welche die Interaktionswahrscheinlichkeit von Transkriptionsfaktoren steuern. Grund ist, daß die Affinitäten, mit denen Regulatoren zu Promotoren hingezogen werden, von diesen Sequenzeigenschaften abhängen.

Contents

List of Figures

List of Tables

1 Motivation

The “circadian” clock is an internal oscillator that keeps track of geophysical time and synchronizes physiological functions, even in the absence of light-dark-rhythms, within around 24 hours, giving reason for its name (circa = about, dies = day, Dunlap [1999]). In mammals, a master clock sitting in the suprachiasmatic nucleus of the hypothalamus orchestrates timing of peripheral clocks, whose existence has now been shown for many organs and tissues (reviewed in Mohawk et al. [2012]). One effect on the immune system was described in 1960 by Halberg et al. [1960]: The authors exposed mice to *Escherichia coli* endotoxin at different times of the day. Their susceptibility to this infection varied ”predictably and significantly along the 24-hour time scale“, because animals died more often when infected during the day with a dose that was not lethal when administered at the middle of the night. Based on knowledge of the clock’s mechanism, Liu et al. [2006b] suggested in 2006 an important regulatory role for PER2 in natural killer cell function based on the observation, that ”*Per2*-deficient mice were more resistant to lipopolysaccharide (LPS)-induced endotoxic shock than control wild-type mice“. A link between the circadian clock system and innate immune responses is also supported by Keller et al. [2009], who demonstrated ”that circadian cytokine response upon LPS stimulation does not depend on circadian cortisol levels but is likely due to functional circadian clocks within immune cells.“ However, the link between the circadian rhythm and timed immune answers is not understood on the molecular level.

To investigate how physiological functions are modulated by the clock system, gene expression was studied globally in many tissues using microarray timeseries during past years. About 10% of genes show daily cycling expression levels under constantly dark conditions. A comparison between circadian genes of liver and heart found ”very few“ common circadian genes, while the distributions of peak times among circadian genes varied markedly among the two tissues (Storch et al. [2002]). Moreover, when Liu et al. [2006a] reanalyzed these data, they found for 20% of the common circadian genes significant phase differences between their expression patterns in liver and heart. The tissue-specificity of phase regulation was even more pronounced in the study of Yan et al. [2008], who compared expression patterns of circadian genes in 14 tissues. They found consistent phases of circadian oscillating genes across tissues only, if the genes were rhythmic in eight or more tissues. In an attempt to explain these phenomena, Masri and Sassone-Corsi [2010] suggested that tissue-specific transcription factors interact with core clock effectors making the composition of gene expression regulating protein complexes dependent on space and time. Furthermore the authors argue for an interplay between metabolism and clock based on the direct regulation of an enzyme critical in the NAD^+ salvage pathway by the main clock transcription factor CLOCK:BMAL1. The metabolic state of a cell seems to be reflected in its chromatin conformation, which can be actively

remodeled under energy consumption. How could one identify the DNA binding proteins responsible for tissue-specific circadian gene expression and phasing?

A canonical method to predict transcription factors that regulate a number of similarly expressed genes is to analyze the sequences near their transcription start sites in comparison to those of genes with different expression properties. Transcription factors involved in the regulation of a gene's transcription bind its promoter region at specific binding motifs. Computational methods are able to detect the enrichment of such motifs in sets of genes responsive to the corresponding transcription factor (Meng et al. [2010]) or in coregulated genes (Kiełbasa et al. [2010]). After this method successfully predicted tissue-specific transcription factors, researchers started to apply it to the search for circadian regulators (Ueda et al. [2005], Bozek et al. [2007], Yan et al. [2008], Bozek et al. [2009], Bozek et al. [2010]). However, they compared promoters of circadian genes against those of all other genes documented in an arbitrary gene database. In view of the tissue-specificity of rhythm and peak phase regulation, the choice of this control group is questionable. In addition, sequence properties of promoters influence the mode of their regulation (Cairns [2009], Valen and Sandelin [2011]). That's why it was suggested to consider two promoter types separately in overrepresentation studies (Roider et al. [2009a], Landolin et al. [2010]).

This study set out to improve our understanding of the link between the immune system and the circadian clock on the transcriptional level based on the microarray data collected by Keller et al. [2009] from mouse peritoneal macrophages. Which transcription factors deliver the timing information of the circadian oscillator to macrophage specific gene expression patterns? In order to assess the cell type specificity of rhythmicity and phase of macrophage circadian genes, a second dataset with timeseries on gene expression in mouse liver cells is called in for comparison (Hughes et al. [2009]). It is assumed, that similar cell functions require the transcription of common genes, while cell type-specific gene transcription is the base for cell type-specific functions. Mouse macrophages and liver cells both contain circadian peripheral oscillators and are involved in the immune system, but nevertheless their specific physiological functions are quite different: Macrophages are major players in the organism's non-specific immune defense: they engulf and obliterate parasites and microbes by phagocytosis (innate immunity). Furthermore they are involved in the activation of lymphocytes by presenting antigens on their surface (adaptive immunity, Elhelu [1983]). As a very versatile cell type they support several other tissue's functions as resident macrophages, e.g. as Kupffer cells in liver (Hume [2012]). The liver contributes to immunity by detoxifying the blood. Besides that, it controls glucose homeostasis, stores vitamins and iron, breaks down hemoglobin and several hormones and converts ammonia to urea (Krucik [2013]).

Considerable effort is invested into the choice of gene sets for proper comparison. The question of tissue- and phase-specific circadian expression regulation is split in two: (1) which transcription factors regulate tissue-specific gene expression at a given time and (2) which transcription factors modulate expression timing so that oscillation occurs? Based on measured timeseries in both cell types all genes detectable by the two different microarray platforms used in the gene expression studies are grouped with respect to their tissue-specific expression, oscillation and timing. To find answers to

the separated questions, genes categorized as circadianly expressed are compared to two categories of background: non-expressed genes and genes expressed without circadian profile. Other variable features (timing, promoter properties and gene category in the second cell type) are controlled for their homogeneity within a gene set and among fore- and background sets. This detailed promoter analysis reveals an estonishing flexibility of gene regulation where small differences in transcription factor binding properties result in large differences of transcriptional outcome.

In the following chapter 2 I will introduce how circadian timing mechanisms in the body intertwine with gene expression regulation. Subsequently, I will explain in chapter 3 how promoter analysis helps to get insights into this field. In the following chapter 4 I describe the data, programs and methods I used before I present my results in the chapters 5 to 7 and discuss them in chapter 8.

2 The circadian clock conducts synchrony in the organism

Living in the 24 hour light-dark-rhythm in most parts of our world required many organisms in all branches of the tree of life to develop an appropriate adaptation system which enables an effective energy balance (Dunlap [1999], Doherty and Kay [2010]). An endogenous oscillator with a period of around a day, the circadian clock, measures time internally and tracks external time by entrainment to various cues (*zeitgebers*), e.g. light, temperature and food. It seems that it had evolved by growing cooperativity between several weaker timing mechanisms (Brown et al. [2012]) and intertwined with redox homeostatic mechanisms (Edgar et al. [2012]).

In mammals, this circadian mechanism regulates many body functions that are time-of-day dependent, most obviously the sleep-wake-behavior, also mirrored in the circadian levels of the sleep hormone melatonin (Lewy et al. [1995]). But also glucose homeostasis (Gatfield and Schibler [2008]), renal activity (Bonny et al. [2013]), blood pressure and heart rate (Wang et al. [2008]), cell division (Johnson [2010]), and more observables are controlled by the circadian clock. This implies that timing plays an important role for many processes in the body, suggesting a high vulnerablility of the organism if the synchrony of the circadian clock is destroyed. Indeed, many diseases are linked to failures in clock rhythmicity (Richards and Gumz [2012]), including cancer (Savvidis and Koutsilieris [2012]), familial advanced sleep-phase syndrome (FASPS, Vanselow et al. [2006]), obesity and diabetes (Gale et al. [2011]), inflammatory diseases (Castanon-Cervantes et al. [2010]), cardiovascular problems (Durgan and Young [2010]) and depression (Barnard and Nolan [2008]), the cited reviews being only the tip of an iceberg.

Understanding the coordination of circadian gene regulation is a key goal of chronobiology research. It is known that the clock influences gene and protein activities on many levels, including DNA accessibility, transcription, translation, post-translational modifications as well as mRNA and protein stability. Here the focus lies on transcription regulation.

How does the body keep track of time? Light sensations from the eye are transferred via the retino-hypothalamic tract to a brain area above the optic chiasm, which is called the suprachiasmatic nucleus (SCN). It is the master clock of the body because it determines the period of all body functions and ablation of it results in arhythmia (Ralph et al. [1990]). The SCN contains about 20,000 neurons which are highly connected due to the exchange of neuropeptides (Maywood et al. [2011]), so that they produce a very stable circadian rhythm (Abraham et al. [2010]) in firing rate and gene expression, also as explants. To date, in almost all cell types of the mammalian body peripheral clocks

have been found (Dibner et al. [2010]): Most tissues express a set of clock genes required for gene expression in a circadian fashion. Due to poorer cellular coupling peripheral oscillators appear to depend on SCN-derived signals, such as hormones, innervation, cytokines, body temperature, feeding times and less known cues, to stay synchronized (Buhr and Takahashi [2013]). The timing of body functions results from the combined action of central and peripheral clocks as shown by the example of a conditional liver clock (Kornmann et al. [2007b], Kornmann et al. [2007a]). How these signals are interpreted tissue-specifically for phase-specific gene expression is not well understood.

Therefore this work focuses on the impact of transcription factors on circadian gene regulation in two peripheral cell types. As their activity unfolds at binding to proper binding sites in gene regulatory DNA sequences, the following section 2.1 discusses general properties of DNA sequence architecture that is relevant for tissue-specific as well as circadian regulation of gene transcription. Following that section 2.2 shifts the focus to the interactions of transcription factors in gene regulatory networks.

2.1 Gene regulation occurs at many levels

For the transcription of a gene the promoter region plays a prominent role. It surrounds the transcriptional start site (TSS), but its range is not clearly defined. A small part of it, about 70-80 base pairs near the TSS, is called the core promoter, because it contains specific binding sites for RNA polymerase II and its cofactors (called general transcription factors), which assemble there to form the pre-initiation compex (PIC). This process alone does not suffice to engange gene transcription in eucaryotes; the binding of specific transcription factors (TFs) to special motifs called *cis*-elements in the extended promoter region nearby or more distant regions like enhancers or insulators is needed to activate or repress gene transcription. Their effect depends on their expression level in the cell as well as on the presence of direct or indirect interaction partners. The occurrence of functional transcription factor binding sites (TFBSs) is often conserved across species, indicating the importance of the transcription factor's interaction with the promoter for the gene's expression. Further aspects of how transcription factors regulate gene expression in time and space are thematized in section 2.1.1.

The genome-wide identification of TSSs revealed a distribution of locations for the majority of genes. According to their shape two promoter classes (LCP and HCP) have been constituted. They distinguish by certain sequence properties, namely the content of guanine (G) and cytosine (C) nucleotides as well as the normalized content of Cytosine-Guanine dinucleotides within their promoter sequences (nCpG). Additionally, they differ in the organization of their DNA accessibility as well as the expression breadth of their downstream genes. The details will be explained in section 2.1.2.

For an interaction to take place the transcription factor's binding site must be exposed to the surface of the DNA molecule. Its accessibility is affected by placement and association of histone proteins, around which the DNA is wrapped, as well as their chemical modifications. Furthermore, chemical marks on the DNA and its structural properties influence transcription levels. This will be discussed in section 2.1.3.

These three levels of gene regulation are illustrated in figure 2.1.

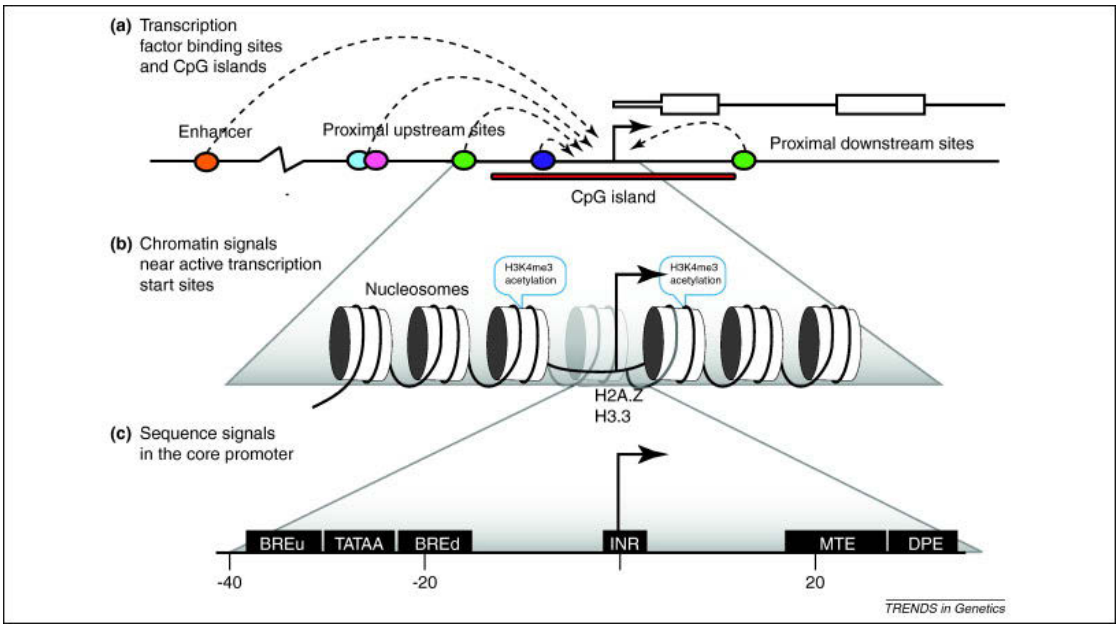


Figure 2.1: **Signals important for transcription activation.**
(a) Transcription factor binding sites and CpG islands. Transcription factors (ellipses) bind to short sequence sections, their binding sites, in the core promoter close to the transcription start site (TSS, arrow) or more distantly at enhancers to activate or repress transcription. Their influence is direct by interacting with the pre-initiation complex or indirect by recruiting enzymes that modify histone tails. **(b)** Chromatin signals around the TSS. White cylinders represent histones with DNA wrapped around, called nucleosomes. While inactive promoters can be covered by canonical nucleosomes limiting access to transcription factor binding sites, active promoters carry activating histone modifications and/or histone variants as indicated. The latter ones occupy the nucleosome free region near the transcription start site and are easier to remove. **(c)** Core promoter sequence patterns. These sequence patterns located within the core promoter have clear positional preferences and are bound by different parts of the pre-initiation complex. Due to their size regions with high content of guanine-cytosine-dinucleotides (CpG islands) that often overlap this region are shown in panel (a). Figure taken from Valen and Sandelin [2011].

2.1.1 The role of transcription factors and their binding sites

Transcription factors are gene regulatory proteins that control gene activity generally, tissue-specifically or in response to environmental, extra- or intracellular signals. They do so by promoting or blocking the recruitment of RNA polymerase to the gene's promoter either directly or indirectly, e.g. by recruiting other transcription factors or modifying hi-

stones. In eucaryotes, usually several transcription factors form a pre-initiation complex to fullfill this task. Its composition varies across tissues and in time "to bestow selectivity of recruitment to chromatin loci corresponding to promoters of clock-controlled genes" (Masri and Sassone-Corsi [2010]).

As distinctive features transcription factors contain a DNA-binding domain (DBD) and a transactivation domain which enable them to bind DNA and activate transcription, respectively. Furthermore, transcription factors often contain a complexation domain that helps them to dimerize with their binding partners (corepressors, coactivators) or sense a cellular signal in form of a ligand. The structure and electrostatic charge of the binding and complexation domains determine their specificity for certain DNA sequences and binding partners. It can change over time by molecular modifications like phosphorylation and acetylation or their reversions, implying dynamic changes in activity and/or protein interactions. Such interaction dynamics have been shown to be important for the regulation of circadian rhythms (Wallach et al. [2013]).

The sites on the DNA where transcription factors bind to are called response elements. In general, they encompass about 6-25 base pairs. Many response elements are located near the transcription start site where the pre-initiation complex forms prior to the start of transcription. Transcription factors bound to distant regulatory elements (enhancers, insulators) interact with other cis-regulatory factors in the promoter region through mediator proteins that stabilize DNA bending to assist the formation of pre-initiation or inhibitory complexes. However, while diffusing through the nucleus, transcription factors interact many times transiently with DNA (Hager et al. [2009]). Although many of such interactions have been viewed as non-functional previously, recent research implies, that additional "decoy" binding sites may influence gene regulation by affecting transcription factor degradation kinetics (Burger et al. [2010]).

The regulation of transcription factor expression is highly tissue-specific and for a certain fraction also phase-specific. The set of transcription factors expressed in a specific cell type is established by a cascade of transcription factors whose identity depends on developmental time and the cellular environment. As many experiments have shown, transcription factors that regulate timing in the cell are expressed in many peripheral tissues as well as in the brain. Nevertheless, additionally to common circadian genes, each tissue contains a different set of circadian genes according to the function of that tissue. The question is, how the few common clock genes regulate the output of different circadian genes among all the tissues. One hypothesis is, that tissue-specific circadian transcription factors are regulated by clock genes and these regulate transcription of targets further downstream in the regulation cascade. This work adresses the question by tissue-specifically analyzing the binding sites of transcription factors overrepresented in circadian genes.

2.1.2 The role of the promoter sequence composition for tissue specificity

In addition to the presence of transcription factor binding sites the sequence composition of a promoter region influences its transcription. Several lines of evidence indicate, that the content of guanine and cytosine base pairs (GC) as well as of cytosine-guanine-

dinucleotides (CpG) in promoters of genes influence their maximum expression level and expression breadth (Vinogradov [2005], Roider et al. [2009a]). This effect may relate to the accessibility of binding sites within these promoters.

As Valen and Sandelin [2011] nicely summarized, it was observed that vertebrate genes with broad TSS distributions often contain CpG islands - these are DNA regions where more cytosine-guanine dinucleotides exist than would be expected given the local ratio of guanine and cytosine nucleotides in the sequence. These promoters with high normalized CpG content (nCpG) are often associated with ubiquitously expressed genes (HCP). On the other hand, promoters with low CpG content correspond to promoters with a sharp TSS distribution and overrepresented TATA boxes in the core promoter, which belong to rather tissue-specific expressed genes (LCP).

Importantly, these two groups of promoters seem to be regulated in ways differing in the priorities of transcription factors and certain chromatin signals as will be further elaborated in section 2.1.3. While binding sites accumulate strongly right upstream the TSS of LCP genes, the proximal region of promoters with high CpG content lack tissue-specific binding signals (Roider et al. [2009a]). The reason for this phenomenon lies in different nucleosome occupancy of the two promoter classes: Genes with low CpG content need TFs to remove nucleosomes covering their TSS while HCPs contain a so-called nucleosome-free region along their TSS-region, where nucleosomes are depleted or only weakly positioned (Valen and Sandelin [2011]). Based on this observation Roider et al. [2009a] suggest to separate these two promoter classes in a promoter analysis assessing the overrepresentation of transcription factor motifs.

2.1.3 The role of chromatin state for DNA accessibility

To fit the long DNA molecule into the nucleus of a living cell, it is tightly wrapped around octamers of histones in sections of 147 base pairs. These complexes are called nucleosomes and cover 75-90% of the DNA. Linkers of about 20-50 base pairs leave space for access of DNA binding proteins like transcription factors and nucleosome remodelers. However, the two complementary DNA strands must be opened for transcription to take place. Hence, the tightness of the interaction between histones and DNA has an impact on transcription activity via remodeling kinetics (Segal and Widom [2009], Padinhateeri and Marko [2011]).

Several modifications of histone tails (phosphorylations, acetylations, sumoylations and methylations) serve as markers to trigger a denser or looser packing of nucleosomes, resulting in hetero- or euchromatin, respectively. By changing the interaction between DNA and histones modifications of the latter proteins affect the process of transcription initiation in the first nucleosomes up- and downstream of the TSS, which cover the promoter region. Histone acetylation is associated with unpacking the chromatin and thus with activating transcription. This is due to the introduction of additional negatively charged residues which repel the DNA's negative phosphate backbone. Histone methylation effectuates the recruitment of other chromatin remodeling factors which can either activate or repress transcription.

Some proteins able to perform these modifications are involved in the control of cir-

cadian rhythms (nicely reviewed in Sahar and Sassone-Corsi [2012] and Ripperger and Merrow [2011]). Equipped with a histone acetyltransferase (HAT) activity the transcription factor CLOCK acetylates its binding partner BMAL1 as well as histone H3 in the promoter region of its target genes (Doi et al. [2006], Hardin and Yu [2006]), which opens the chromatin for transcription. Furthermore, CLOCK interacts with the histone methyltransferase MLL1 and the histone deacetylase (HDAC) SIRT1, which are epigenetic regulators that are able to modify the chromatin according to environmental stimuli, such as nutrient availability (Sahar and Sassone-Corsi [2012]). SIRT1 counterbalances CLOCK's histone acetylation activity in a time-of-day dependent manner, although its mRNA is not rhythmic (Nakahata et al. [2008]). The circadian activity of SIRT1 is based on its dependence on its cofactor NAD^+, whose rhythmic biosynthesis results from gene expression control of a key rate-limiting enzyme, nicotinamide phosphoribosyltransferase (NAMPT), by CLOCK, BMAL1 and SIRT1 (Nakahata et al. [2009]). The cycling of *Bmal1* transcription results from rhythmic recruitment of HDAC3 by the nuclear receptor REV-ERBα and its corepressor (NCoR1) to mediate transcriptional repression (Yin and Lazar [2005]).

The two promoter classes introduced in section 2.1.2 are configured differently with respect to these modifications (Valen and Sandelin [2011]). CpG-rich promoters are easily bound by CFP1/SET1, a complex of a CpG-binding protein with a H3K4-methyltransferase, that introduces three activating methyl groups (me3) to the lysine residue K4 in histone H3. Although H3K4me3 is one of the most common methylation marks at a promoter associated with activation, many promoters carrying it are kept in the "poised" state and do not produce mRNA until - in a second layer of control - RNA polymerase II is released and elongation starts. This explains the often observed nucleosome-free region in the promoter region of HCP genes. LCP genes often lack a nucleosome-free region, since they do not carry activating marks by default. They mainly depend on additional chromatin-remodeling factors (SWI/SNF), which remove the nucleosome covering their transcription start site dependent on the energy supply via ATP (Vignali et al. [2000]). Presumably, they contain more transcription factor binding sites to attract transcription factors who recruit these remodelers tissue-specifically. Repressing histone methylations is a further means of securing downregulation of promoter activity in many tissues, which mainly occurs in LCPs.

Thus, although the classification of LCPs and HCPs is not exclusive, regulation of tissue-specific gene transcription is closely interwoven with the DNA accessibility determined by the promoter properties of the genes. Therefore a key innovation of this study is the separation of LCP and HCP genes in promoter analysis.

2.2 Transcription factor networks determine timing and tissue-specificity

Considering the length of DNA, its strechwise tight packing and the vast majority of unspecific binding sites it seems astonishing, that a transcription factor finds its functional binding sites in the promoter regions of its target genes. By in vivo imaging Hager

et al. [2009] showed, how transcription factors find their targets by three-dimensional scanning of the genome using activities like diffusion, sliding along a DNA strand and hopping over DNA loops. While many transcription factors spend only a short period of time at a particular binding site, some are bound more stably. A DNA stretch to which proteins repeatedly attach might remain accessible, because it is blocked against nucleosome occupation. However, nucleosome remodeling is a dynamic and continuous process, and as transcription factors are competitors of histones for DNA binding, they either require or initiate local chromatin reorganization.

However, a bound transcription factor cannot initiate transcription of a target gene by itself. The interaction of several transcription factors is necessary to recruit RNA polymerase. Current in vivo data support a model in which polymerase and regulator complex assembly occurs by random collision of subunits at the promoter (Hager et al. [2009]). To establish a higher promoter occupancy in spite of transient transcription factor binding to DNA, transcription factors need (1) easy access to their DNA binding sites and (2) to dwell there long enough to have the chance to meet their binding partners.

Pioneering transcription factors care for the first part: They can bind to densely packed chromatin and recruit nucleosome remodelers to open closed chromatin structures and make the DNA accessible for other transcription factors to come (Zaret and Carroll [2011]). Examples for such pioneering transcription factors are the Forkhead box (Fox) family and GATA factors which are expressed in the foregut endoderm and necessary for liver induction and development as well as C/EBPα and C/EBPβ for activating the macrophage program in B cells. However, each nucleosome remodeling needs further transcription factor binding to stabilize the new structure.

The probability to meet binding partners is increased by prolonged residence times of polymerase components at their DNA binding sites (Hager et al. [2009]). This can be achieved by protein interactions or conformational changes. For example, a conformational change of DNA due to protein binding may facilitate another molecule's binding. Its binding may even stabilize the conformational change and likewise the binding of the first protein to DNA, increasing the chance of their interaction.

Hence, tissue-specific chromatin remodeling and binding dynamics of transcription factors play important roles in gene regulation. Interactions between transcription factor proteins and their target genes are described by transcription regulation networks. Their smallest building blocks are called network motifs, which refer to recurring circuits of interactions. Their combination may lead to more complex dynamical gene expression patterns, including oscillations on different time scales (Alon [2007]). One known rhythmic expression mode is regulated by the circadian clock, whose basic mechanism will be discussed next.

2.2.1 The core clock mechanism

Circadian gene expression in mammalian cells relies on interlocked positive and negative feedback loops of transcription and translation called the basic clock mechanism (figure 2.2). The dimerized transcription factors CLOCK and BMAL1 activate the transcription of *period* (*Per1, Per2*) and *cryptochrome* (*Cry1, Cry2*) genes. When trans-

lated, complexed and phosphorylated, these gene products return to the nucleus and inhibit their own transcription by binding to CLOCK and BMAL1 (Dunlap [1999]). A new expression cycle starts after the inhibiting proteins are degraded. Cycling BMAL1-levels are established by the transcription factor REV-ERBα, a target of CLOCK:BMAL1 that inhibits *Bmal1* expression. Genes which are involved in the described mechanisms are called "core clock genes" from now on.

By contrast, "clock controlled genes" constitute the output of clock regulation: they also show a circadian rhythm in their gene expression pattern, but do not feed back directly to the clock mechanism. Furthermore, they are specifically expressed or regulated in peripheral tissues. On the question of how the clock mechanism mediates tissue- and phase-specificity to circadian genes, Balsalobre suggested that "one common core oscillator in peripheral cells may regulate many cell-specific clock controlled genes (*ccgs*) by directly regulating a relatively small number of tissue-specific circadian transcription factors that would then regulate a plethora of target *ccgs*" (Balsalobre [2002]). However, these regulation cascades remain to be explored. On this journey the "comparison of circadian oscillating genes and their oscillating patterns across different tissues" can be helpful to understand the tissue-specific functions of circadian rhythm (Yan et al. [2008]).

2.2.2 Transcription factor interactions direct tissue-specificity

In the very simple picture drawn until now, for each cell type exists a set of transcription factors which regulate the cell type's specific gene expression. As shown by Ravasi et al. [2010], reality is more complicated. The authors propose that not a set of expressed transcription factors, but their interactions determine tissue-specificity. In this context, an interaction is very generally defined as the co-expression and co-localization of transcription factors. The authors observed, that "TFs with few interactions tend to be expressed in a tissue-specific pattern while TFs with many interactions - so called network 'hubs' - tend to be expressed across many tissues". In their model, tissue-restricted TFs (specifiers) interact with broadly-expressed TFs "increasing the number of possible combinatorial events only in certain tissues or during tightly-regulated developmental processes". In agreement with this, Nowick and Stubbs [2010] describe gene regulatory networks as hierarchical organizations, in which the transcription factors are in principle interchangable, but their network structure is crucial for functional gene regulation.

With regard to transcription factor binding to promoters of genes the "fuzzy puzzle" model presented by Kel et al. [2000] illustrates this hypothesis graphically (figure 2.3): "The structure of regulatory sequences on one hand and the specific features of transcription factors on the other hand provide a possibility to encode several regulatory programs within one regulatory region. It is known that each transcription factor has the ability to bind to a variety of different DNA sites. This is maintained by flexible mechanisms of DNA-protein interactions, when DNA conformation rather than the particular sequence context often play the major role in selection of DNA targets. In addition, the ability of TFs to operate through a so-called induced fit mechanism (when a TF becomes finally structured only upon interaction with DNA) greatly relaxes the

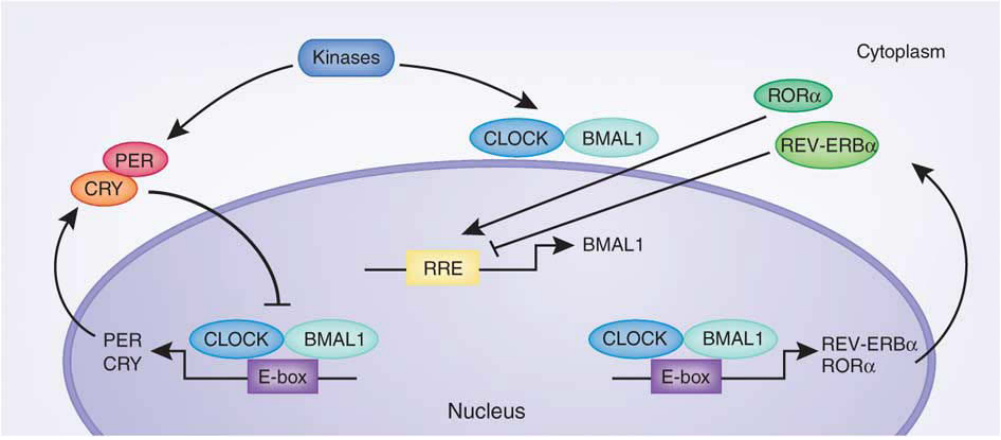


Figure 2.2: **The circadian clock network.**
The core circadian transcription factors, CLOCK and BMAL1, direct Ebox-mediated transcription of clock-controlled genes, including acticators and repressors of the circadian system. PER and CRY protein translation occurs at night and subsequently causes repression of the core CLOCK:BMAL1 transcriptional complex. Degradation of the repressors PER and CRY prompts a new circadian cycle whereby CLOCK:BMAL1 transcription is reinitiated. In addition to transcriptional regulation, post-translational modifications are crucial for the modulation of circadian proteins. The figure shows only phosphorylation, which acn be elicited by sveral kinases, including CKIϵ, CKIδ, CK2α, GSK3β and AMPK. Other post-translational modifications of clock proteins include acetylation, sumoylation and ubiquitination. RRE, REV-ERB/ROR response element. Figure taken from Masri and Sassone-Corsi [2010].

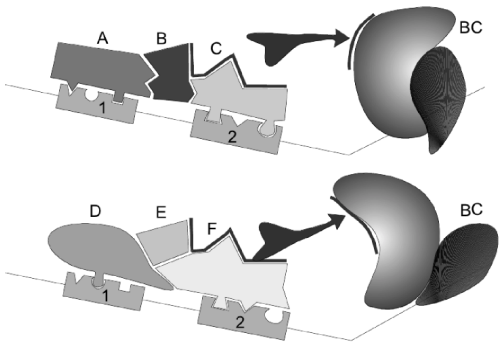


Figure 2.3:
The fuzzy puzzle hypothesis.
Multiple regulatory messages can be encoded within the same regulatory region due to the structure of regulatory sequences on one hand and the specific features of transcription factors in the other hand. A,B,C and D,E,F –two sets of transcription factors; 1,2 –two sites in DNA; BC –basal complex. Figure taken from: Kel et al. [2000].

restrictions from binding to various DNA sites. Besides that, the protein-protein interactions between different transcription factors in the multiprotein regulatory complex become very important. Protein-protein interactions could stabilize some low-energy protein-DNA contacts thus additionally widen the variety of target sites for particular

transcription factors. The huge diversity of transcription factors functioning in the living cells multiplied by the wide choice of target sites for each TF give rise to a precondition to form multiple alternative DNA-protein complexes on the same gene regulatory region. As a result extremely complex patterns of gene expression are observed.“

In conclusion, transcription factor interactions are important for transcription activation and depend on tissue and time due to differential and permanently changing DNA conformations.

2.2.3 Transcription factors in a phase vector model

Based on a promoter analysis of eight circadian genes common to six tissues Yamamoto et al. [2004] found that the peak timing order of the circadian genes related to the presence of binding sites within their promoter regions which were bound by specific transcription factors involved in the core clock mechanism. The authors proposed that ”cyclic timing of all clock and clock controlled genes may be dependent on several transcriptional elements including three known elements, EBox, RORE and DBPE“. This model was further developed by Ukai-Tadenuma et al. [2011], who used a phase vector model to predict a gene's peak phase based on the transcription factor binding sites present in its promoter region. This view on circadian peakphase regulation underlines the model in which transcription factor interactions determine the outcome of gene expression.

3 How promoter analysis provides insights into circadian gene regulation

As transcription factors mainly influence the expression level by binding to certain motifs in the promoter region of target genes, important aspects to understand gene regulation are how transcription factors bind to DNA, how they find their specific binding sites and which transcription factor binds to which DNA sequence motif. Modern technologies helped to accumulate answers to these questions, which can in turn be used to predict regulators of genes with common expression patterns based on their promoter sequences. After a short survey on the classifications of transcription factors and their binding sites (section 3.1), the next sections will be dedicated to introduce the method of promoter analysis (section 3.2), summarize previous findings by overrepresentation analyses of promoters (section 3.3), to spot unanswered aspects and describe this work's aim (section 3.4).

3.1 Characterizing transcription factor binding to promoters

Transcription factors can be functionally classified based on their signal responsiveness and localization or based on their three-dimensional structure. The latter possibility facilitates the comparison of their DNA binding preferences. Accordingly, transcription factors are sorted into five superclasses (Stegmaier et al. [2004]), named by the structural features of their DNA-binding domains: (1) basic domains, (2) zinc-coordinating domains, (3) helix-turn-helix domains, (4) β-scaffold factors with minor groove contact and (5) other transcription factors. Each class contains families of transcription factors with similar but distinct binding site preferences (Stegmaier et al. [2013]).

To detect binding sites, to which transcription factors bind, antibodies specific for the transcription factor in question are employed in a Chromatin ImmunoPrecipitation (ChIP) experiment. After covalent crosslinking all DNA associated proteins to the site of their attachment, fragmentation of the DNA, immunoprecipitation of the transcription factor of interest along with its bound strech of DNA using specific antibodies against it, the crosslinking is reversed and the DNA sequences are followed by sequencing. The alignment of many binding site sequences leads to a tabular description of the binding site, called positional count matrix, showing the total counts of each of the four nucleotides for every position of the motif. Databases like TRANSFAC (Matys et al. [2006]), JASPAR (Sandelin et al. [2004]) and SwissRegulon (Pachkov et al. [2007]) store hundreds of such motif descriptions.

The binding sites found by ChIP experiments have been verified by measuring the binding strengths of transcription factors to protein binding arrays, which are microar-

rays spotted with all combinations of oligonucleotides of 10 base pairs length (Badis et al. [2009]). This study highlights the fact that some transcription factors even recognize secondary motifs slightly different from the primary motif.

3.2 Promoter analysis

Using the described transcription factor motifs, promoter analysis reverses the process of characterizing transcription factor binding preferences and tries to predict regulators of genes based on the binding sites found in their promoter sequences. While the concentration of free transcription factors and the number of accessible specific binding sites vary among cell types and with time, the promoter sequence of genes provides constant information. Together with the regulators' binding profiles it can be used to estimate the probability of transcription factor binding. This is based on a score that characterizes the similarity of each possible binding site in the DNA relative to a particular transcription factor's most favourite motif. By scanning the chosen promoter region in this way, putative binding sites can be predicted.

Let's look at this in more detail. In a first step the positional count matrix for the experimentally determined binding motif is recalculated to a positional frequency matrix, which gives the ratio of each nucleotide at each position compared to the total amount of captured sequences. Using information theory (Shannon [1997]), the frequencies of nucleotides in the motif and in an appropriately chosen background model describing general nucleotide frequencies in the promoter region are combined in the calculation of log likelihoods yielding a positional weight matrix (PWM). The negative logarithm of this ratio is called a weight and describes for each nucleotide at each position in the motif the information content of this base with regard to the transcription factor binding site of interest. The weights for the occurring nucleotides at each position in an arbitrary short promoter segment sum up to the above mentioned score. Thus, the consensus sequence of the motif has the highest score.

To characterize the binding of a transcription factor to a promoter several strategies have been developed (figure 3.1). One possibility is to count the number of binding sites occurring in the sequence which score better than a certain level (hit based method, Rahmann et al. [2003]). By using a threshold the question of how to evaluate the different qualities of binding sites is reduced to a binary problem, and the number of positive answers depends on the chosen cutoff, which might also differ among various motifs. Another way is to sum the scores of all sites in a weighted manner based on a biophysical model of the binding energies between transcription factor and DNA. According to the Boltzmann distribution this model assumes, that binding sites with larger scores have more chances to be bound by a transcription factor than the ones with lower scores. Taking such a model into account, the affinity of a transcription factor to a certain promoter region is calculated as the expected number of binding regulators. While this TRAP method (TRanscription factor Affinity Prediction, Roider et al. [2007]) circumvents the use of a threshold, it cannot locate the site of most probable binding (Thomas-Chollier et al. [2011]). It is also noteworthy, that the affinity does not

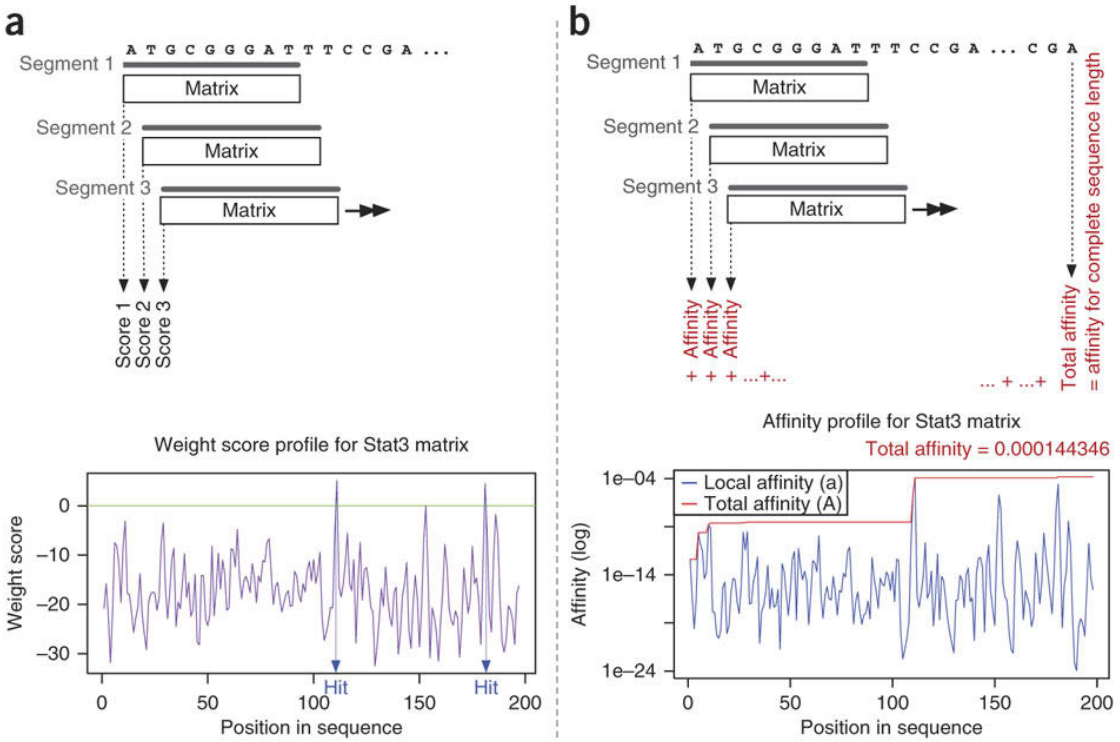


Figure 3.1: **Methods for prediction of transcription factor binding.**
(a) Hit-based method. Alignimg the matrix to the sequence, a score is given to each segment. Only the positions of scores reaching the predefined threshold (*green line* for an arbitrary value of 0) are reported as TFBS hits (*blue arrows*). **(b)** Affinity-based method as implemented in TRAP. Aligning the matrix to the sequence, the total affinity value is obtained as the sum of all affinity values for each segment. The cumulative TRAP score is shown in *red*. Figure taken from Thomas-Chollier et al. [2011].

express an observable strength of transcription factor binding. Other methods for TFBS prediction can be associated with these main strategies.

When analyzing large sequences for the presence of transcription factor binding sites many hits will be due to chance. Unfortunately, their functionality cannot be predicted, and many false positive hits will be reported. Approaches to narrow down the number of putative hits include (1) to check for binding sites only in relevant promoter regions near the transcription start site, (2) to reevaluate the hits based on phylogenetic footprinting, which accounts for conservation of sites in orthologous promoters of related species, and (3) to analyze a group of co-regulated genes to see whether their promoters are enriched with a certain binding site which might cause their co-regulation.

Concerning the last point, Meng et al. [2010] has shown that promoters of differentially regulated target genes of knocked out transcription factors have a higher affinity to this regulator than promoters of other genes. Likewise, Roider et al. [2009b] was

able to confirm known associations between tissues and transcription factors based on promoter affinities of tissue-specific genes to these transcription factors. Thus, the overrepresentation of a motif hints at a functional part of the corresponding regulator. This perception was the basis for overrepresentation studies in the past (Bozek et al. [2007], Bozek et al. [2009], Bozek et al. [2010], Yan et al. [2008]) as it is for the analyses conducted for and presented in this thesis.

3.3 Previous findings on circadian gene regulation based on promoter analysis

According to the described core clock mechanism, target genes of the transcription factors CLOCK:BMAL1 share a common feature in their promoter sequences which contains the binding site of these transcriptional activators. This feature is termed Ebox and is characterized by the canonical consensus sequence "CACGTG". In fact, it has been found overrepresented in circadian genes common to many tissues compared to all other genes in the database (Bozek et al. [2009], Yan et al. [2008]). Furthermore, experiments relying on chromatin immunoprecipitation combined with deep sequencing in liver showed BMAL1 binding around zeitgeber time 6 at tandem Eboxes, consisting of two Eboxes separated by a linker (Rey et al. [2011]). However, only 60% of all detected binding locations were bound in a significantly rhythmic manner. This proportion increased in subsets of binding locations with higher binding strengths. Therefore, the rythmic transcription of a gene does not solely depend on the presence of an Ebox. Suggestions on further determinants of the cycling occupation of an Ebox include: (1) flanking regions of the motif, (2) the promoter structure, in which the motif is embedded (Munoz et al. [2006]), or (3) additional *cis*-elements for co-binding factors (Ueshima et al. [2012]). However, many other transcription factors bind a more general motif similar to an Ebox: "CANNTG" (Massari and Murre [2000]). When trying to determine a gene's expression timing, it should be considered, that competition and cooperation among these factors and with their interacting proteins depend on (1) the cellular concentration ratios of the present regulators and (2) the genomic context for transcription factor binding at the specific promoter.

Besides the Ebox, other motifs have been found overrepresented in circadian genes. Yamamoto et al. [2004] ordered genes by the peak phase of their rhythmical expression patterns and found that their timing depends on several transcriptional elements, including Ebox, Dbox and RORE. The phase of expression results from the combined utilization of these motifs within the promoter regions of circadian genes (Ueda et al. [2005]): (1) Ebox motifs as binding sites for the transcription factors CLOCK and BMAL1, (2) RORE motifs as binding sites for competing ROR and REV-ERB family members, which activate or repress transcription, respectively, and (3) Dboxes as binding sites for PAR bZip factors like the activators DBP, HLF, TEF and the repressor E4BP4. This model was further elaborated to a vector model by Ukai-Tadenuma et al. [2011] allowing to predict the phase of the highest expression by a combination of these elements.

After Storch et al. [2002] addressed the small extent of overlap between circadian genes

expressed in liver and heart, Yan et al. [2008] compared circadian genes among fourteen tissues and found less common overlap the more tissues were compared. With respect to peak phases of circadian genes in different tissues two observations stand out: (1) The peak phase distributions of two sets of tissue-specific genes differ (Storch et al. [2002]) and (2) the tissue-specific peak phases of individual genes are the more similar in the more tissues the genes are expressed rhythmically (Yan et al. [2008]). The top five phase-specific transcription factor families are determined as binders to Ebox, AP-2, CRE, SP1, and EGR motifs (Yan et al. [2008]). However, the associated circadian phases of these TF families among different tissues vary considerably. The authors suggest, that "the gene regulatory network responsible for generating spatial expression variation across tissues may be also responsible for generating the temporal expression variation."

Since the amplitude or phase of some clock controlled genes differ between tissues, there must be an additional tissue-specific regulatory part in place. Using annotated affinities (PASTAA), Roider et al. [2009b] were able to show that many transcription factors act tissue-specifically. Which of these are also involved in circadian regulation? A meta-analysis carried out by Bozek et al. [2009] predicted transcription factors involved in phase- and tissue-specific circadian regulation based on binding site overrepresentation analysis. To compensate for the higher GC content of circadian foreground genes compared to all other genes in the database, they employed a GC-matched backgrond model in their overrepresentation analysis. Besides known regulators (CLOCK:BMAL1, DBP, HLF, E4BP4, CREB, RORα) they identified recently described ones (HSF1, STAT3, SP1 and HNF-4α) as well as new candidates: PAX-4, C/EBP, EVI-1, IRF, E2F, AP-1, HIF-1 and NF-Y. One of their promising candidates, NF-Y, has recently been validated to functionally regulate *Bmal1* transcription (Xiao et al. [2013]). This work aims at a more detailed overrepresentation analysis comparing regulatory influences on circadian genes in mouse macrophages and liver cells.

3.4 Aim of the study

Lacking a collection of non-circadian genes, previous analyses compared promoters of circadian genes of any tissue to all other promoters annotated in the database. When asking for tissue-specific regulation this approach skews the analysis by mixing up related, but distinct questions: (1) Which transcription factors mediate tissue-specific gene expression? (2) Which transcription factors support circadian oscillation of gene expression? (3) Which transcription factors exhibit more influence on the expression of circadian genes than of non-circadian genes? To deskew these questions in promoter analysis, genes should be selected more carefully for comparison. This includes grouping genes for several criteria, that can be regulated tissue-specifically (expression, rhythmicity and peak phase) or may be used for tissue-specific regulation (promoter properties). To analyze tissue-specific differences between circadian and other genes, proper foreground and background sets need to be chosen, whose elements differ in one, but are as similar as possible with respect to the other criteria. An analysis like this could help to answer the question how one particular gene can be expressed rhythmically in one cell

type while it remains non-oscillating in an other cell type.

The fact, that the Ebox "CACGTG" was found overrepresented in circadian genes, a subset of expressed genes, compared to a majority of non-expressed genes in the database, could be interpreted in a different way. Due to the packaging of DNA, its occupancy with nucleosomes and epigenetic modulation many binding sites in non-expressed genes are not accessible. However, pioneering transcription factors can recruit factors to change these modes (Zaret and Carroll [2011]). Hence, the CLOCK:BMAL1 heterodimer could have pioneering properties leading to rhythmic chromatin remodeling. In fact, activating marks on promoters of clock controlled genes are modified in a rhythmic manner by chromatin modifying enzymes (reviewed in Bellet and Sassone-Corsi [2010], Aguilar-Arnal and Sassone-Corsi [2013]). However, it is conceivable, that transcription factors regulating circadian genes also influence other genes. If the Ebox served as binding site for pioneering factors it could be found overrepresented in non-circadianly expressed genes even when comparing them to non-expressed genes. It would then be even more difficult to detect overrepresentation of Eboxes in circadian genes compared to non-circadian genes, if both sets were to be subsets of expressed genes. On the other hand, the relatively small number of common circadian genes among different tissues argues for a more specific control of CLOCK:BMAL1 transactivation. Alternatively, the Ebox binding site could be sequentially bound by different transcription factors, as it is known that many other tissue-specific proteins bind to an Ebox of the more general consensus "CANNGT" (e.g. Massari and Murre [2000], Munoz and Baler [2003], Adhikary and Eilers [2005]). The proposed analysis of tissue-specific gene groups could be a useful tool to address the question for additional tissue-specific sequence features that tissue-specifically distinguish circadian from other genes.

To understand the tissue-specific and precise timing of circadian genes is an intriguing task. Experience has shown that the timing of a common circadian gene is the more similar between tissues the more cell types express it rhythmically with a period of about 24 hours (Yan et al. [2008]). Moreover, genes with broad expression usually have a higher GC and nCpG content than rather tissue specific genes (Roider et al. [2009a]) and promoters of circadian genes have a higher GC content than all other genes in the database (Bozek et al. [2009]). This implies that general promoter properties like GC and normalized CpG content play a role in tissue-specific circadian gene regulation. It is therefore advisable to include them as criterium for gene grouping in promoter analysis as suggested by Roider et al. [2009a].

A second aspect for tissue-specific circadian timing are transcription factor interactions. Obviously the absolute affinities of transcription factors to certain gene promoters are the same in each cell type, because they depend only on the promoter sequences and the regulators' binding preferences, which are constant. But the environment differs between tissues with regard to (1) the presence and concentrations of the transcription factors and their competing or auxiliary factors as well as (2) the chromatin state, making different sets of binding sites accessible. Tissue specific gene expression is managed by interaction of so-called specifier (tissue specifically expressed) and facilitator (more broadly expressed) transcription factors (Ravasi et al. [2010]). The interactions of such transcription factors with core clock regulators might influence the cell type's specific

timing of gene expression (Masri and Sassone-Corsi [2010]). To capture binding sites for these interacting partners it is important to compare circadian and other genes based on their expression timing. This necessitates to consider phase groups in promoter analysis.

A last decision prior to promoter analysis concerns the invariable which should be compared between circadian and other genes. Experiments targeting BMAL1 in liver cells using chromatin immunoprecipitation and deep sequencing (ChIP-Seq) revealed that the proportion of circadian bound BMAL1 targets increases with BMAL1 binding strength (Rey et al. [2011]). However, measured transcription dynamics showed that transcription factors diffuse in three dimensions through the cell. The speed of their meandering is accelerated by transient DNA binding (Hager et al. [2009]). Therefore, binding strength can be approximated best by the predicted TF-promoter affinity that adds the impact of each possible binding site based on a biophysical model (Roider et al. [2007], Manke et al. [2008]). With this in mind, tissue-specifically regulated circadian genes would be expected to bind BMAL1 stronger than common non-circadian genes, while common circadian genes would be among the strongest binders.

The ideas presented here are used to define suitable background sets for circadian gene subgroups and their separate comparisons to non-expressed as well as non-circadian genes. More specifically, background genes are chosen with regard to the tissue- and phase-specificity and to match GC and CpG properties of foreground genes. With this refined method this work aims to predict transcription factors important in tissue- and phase-specific circadian gene expression regulation based on overrepresentation of their binding sites.

4 Materials and methods

4.1 Experimental data sources

The analysis is based on time series data of gene expression in mouse peritoneal macrophages (Keller et al. [2009]) and liver cells (Hughes et al. [2009]). Over two consecutive days mRNA levels were measured using Affymetrix' GeneChip mouse gene 1.0ST Array (35557 probesets) or Mouse Genome 430 2.0 Array (45101 probesets) respectively. Different intervals of measurements yielded 12 timepoints for macrophages (every 4 hours) and 48 timepoints in liver cells (hourly).

For the macrophage dataset, microarray fluorescence measurements of probe spots were normalized by Dr. Kuban, Laboratory of Functional Genome Research at the Charité Core Facility (LFGC). The liver dataset was downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). For further analysis the fluorescence measurements were $\log_2$-transformed.

4.2 Databases

Sequences Promoter sequences were downloaded from the Ensembl version 58 (mus musculus NCBIM37, www.ensembl.org) containing 35958 gene identifiers. Transcription start sites were defined as the 5' end of an Ensembl gene, as annotated in the Ensembl database. Downloaded sequences encompassed symmetrical regions around the transcription start sites (TSS) of 1000 base pairs, since it is known that many TFBSs cluster closely around the TSS (Consortium [2007]).

Transcription factor classes The assignment of transcription factors to families based on their DNA binding domains is taken from Stegmaier et al. [2004]. Information on some motifs missing in their list was completed if possible from literature: for two STAF binding sites from Schaub et al. [2000], for the WHN binding site from Schlake et al. [1997], for the MEF3 and PTF1_β motifs from Crown Human Genome Center [2013].

Binding motifs Tabular descriptions of transcription factor binding sites were taken from the TRANSFAC database version 12.1 (Matys et al. [2006]). It contains 610 positional count matrices representing vertebrate binding motifs that are putatively bound by around 1500 transcription factors.

The positional count matrix for the double Ebox motif was taken from Rey et al. [2011].

4.3 Programs

For clear probeset mapping to genes Ensembl and Affymetrix annotations were used and processed using the programming language Perl (v5.10.1).

To fit a sine wave to the measured time series and calculate the p-value for the fit's significance compared to a constant fit the program CircWave 3.3 by Roelof A. Hut and Leon Steyvers was used (Hut and Steyvers [2007]).

Transcription factor affinities were calculated using the program ANNOTATE, a new version of TRAP (Roider et al. [2007]), which is regularly updated by Morgane Thomas-Chollier (Thomas-Chollier et al. [2011]).

Gene group definition, phase grouping, background matching, overrepresentation analysis, statistical computations, tables and graphics were done using R version 2.15.2 (R Core Team [2012], Bembom [1990]).

4.4 Statistics

4.4.1 Binding site prediction in promoters

Transcription factor binding sites are described in a tabular format, in which for each position of the binding site the frequencies of the four possible nucleotides are listed. How well a short site on a promoter sequence matches this tabular binding site description (motif) is evaluates by a score, that is the sum of position-dependent weights for each nucleotide in the binding site's sequence. To obtain the weights, the observed transcription factor binding preference documented in the motif matrix is compared to the randomly expected nucleotide frequency at each single position within the motif:

$$\text{weight} = \log_2 \frac{\text{observed frequency}}{\text{expected frequency by chance}} \tag{4.1}$$

The expected base frequency is calculated from a background model describing the base pair distribution in a random sequence with the same GC content as the promoter of interest. Thus, if the putative binding site contains a certain nucleotide at a particular position that has higher (or lower) frequency in the motif than it would be randomly expected, it gets a positive (or negative) weight.

4.4.2 Hierarchical multiple hypothesis testing

Hypothesis testing is used to assess wether an observation can be explained by chance alone. One tries to falsify the assumption of chance as the causative agent for the observation and therefore formulates it as the null hypothesis (H_0). If the calculated probability to get an observation at least as extreme as the one at hand falls bellow a predetermined significance threshold α, the null hypothesis is rejected. This test decision may be erroneous. Two types of errors exist: (1) If an observation that really occured by chance is assumed to be a sign of different conditions, it is a type I or α-error, termed

false positive (FP). (2) If the observation came up due to changed conditions that were not recognized by the test, it is a type II or β-error, termed false negative (FN).

The multiple testing problem The error of a false positive detection in one hypothesis is given by the chosen threshold α. However, the error of at least one false detection in m independent hypothesis tests is $1-(1-\alpha)^m$. It converges to 1 with growing number m of tests. Therefore, single p-values below the threshold α do not correspond to significant results in a study applying multiple tests.

False discovery rate To solve this problem, several methods have been used so far; in this work I refer to the False Discovery Rate (FDR) introduced by Benjamini and Hochberg [1995]. It is defined as the proportion of false discoveries (FP) among all discoveries (D). Accordingly, all p-values (with N being their number) - sorted by their value and indexed with i - can be adjusted to control the false discovery rate at threshold $q_i = p_i N/i$ by enforcing monotonicity: $q^*{}_i = q_i$ with $q_i = min(q_k)$ and $k \geq i$. This method is implemented in R under the function `p.adjust()` with the specified method "BH" (Yekutieli and Benjamini [1999]).

Hierarchical testing A set of p-values intended for BH-adjustment is assumed to be collected from identical experimental conditions. However, in this work, gene sets with distinct information on phase- and tissue-specificity of their expression timing and rhythmicity are analyzed. Because hypotheses of different subgroups (families) are not interchangeable, their p-values cannot be pooled together to determine the FDR (Efron [2008]). Instead, hypotheses are grouped into a hierarchical tree of families as described in Sankaran [2011]. The FDR of hypotheses within a family is controlled as previously described (BH). After that, the following formula helps to determine the proportion of false disoveries in several families of multiple hypotheses (Yekutieli et al. [2006]):

$$FDR_\alpha \approx \alpha * \frac{\text{Total discoveries} + \text{Number of families}}{\text{Total discoveries} + 1} \tag{4.2}$$

As long as the number of rejected hypotheses exceeds the number of families, the FDR will be controlled close to the level α used in testing individual families.

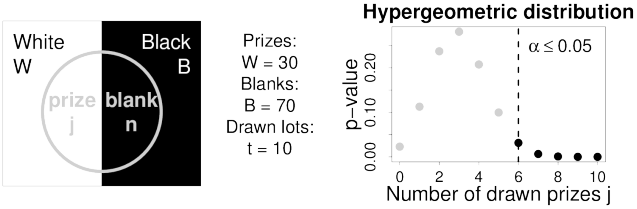
4.4.3 Wilcoxon test

Affinity distributions of promoters in a circadian foreground set and a 500 times larger background set containing randomly sampled non-circadian or non-expressed genes are compared using the non-parametric one-sided Wilcoxon rank sum test. After ranking all affinities together, it calculates a test statistic from the sum of foreground and background ranks to judge by its size on the significance of distribution differences. Due to the sensitivity of the test, it detects whether the medians of the two affinity distributions differ and the circadian affinites are shifted to higher values.

4.4.4 Hypergeometric test and Fisher's exact test

A hypergeometric test calculates the probability for the size of an overlap between two subsets chosen from a larger group of elements. Let us assume, I blindly draw a number of lots (t) from an urn which contains B blanks (black balls) and W prizes (white balls) as illustrated below. With which probability p do I hold $j = t - n$ prizes? This is given by the formula shown below. The probability for the event to draw j or more prizes to occur by chance is equal to the sum over the hypergeometric p-values for all possible events to draw j or more prizes (up to the minimum of the number t of drawn lots or the number W of prizes available). This is equal to the p-value of the one-sided Fisher's exact test, and if it falls below the threshold α chance as cause is rejected.

$$p = \frac{\binom{W}{j}\binom{B}{t-j}}{\binom{W+B}{t}} \tag{4.3}$$



4.4.5 Kolmogorov-Smirnov-test and QQ-plot

To evaluate whether the GC and nCpG content distributions among foreground and background sets are similar, the Kolmogorov-Smirnov-test is applied. Between two independent samples it records any difference in the shape of measured value distributions by comparing their relative cumulative curves. If their maximal difference exceeds a test threshold, it can be assumed that the two distributions differ significantly.

The quantile-quantile-plot (QQ-plot) nicely illustrates such differences as deviation from the bisecting line in the first quadrant of the Cartesian coordinate system. Therefore, the quantiles of the background sets (y-axis) are plotted against the quantiles of the foreground set (x-axis).

4.4.6 Receiver operating characteristic and area under the curve

An observed daily-rhythmic (circadian) expression pattern may be caused by gene regulation, but it may also have occured by chance. Therefore it must be possible to estimate the false positive rate for each rhythm detection method to give a significance statement. In this work this was done empirically by analyzing random profiles. They were created as permutations of the measured time series.

To illustrate the ability of a binary criterium to classify between circadian and non-circadian groups, Receiver Operating Characteristics (ROC) are calculated and shown as ROC curves. Therefore, measured (M) and randomized (R) timeseries are used to estimate the number of correctly and erroneously detected profiles as the threshold of the classifier is varied. Then sensitivity (true positive rate, $\frac{M_{pos}}{M_{pos}+M_{neg}}$) is plotted against $1-$specificity (false positive rate, $\frac{R_{pos}}{R_{pos}+R_{neg}}$). The area under the ROC curve is a measure for the discrimination strength of the classifier used.

5 Subsets of circadian genes reveal their heterogenous regulation

This chapter introduces the main concepts of the study and brings them into context within the workflow (figure 5.1). More detailed results are presented in chapters 6 and 7.

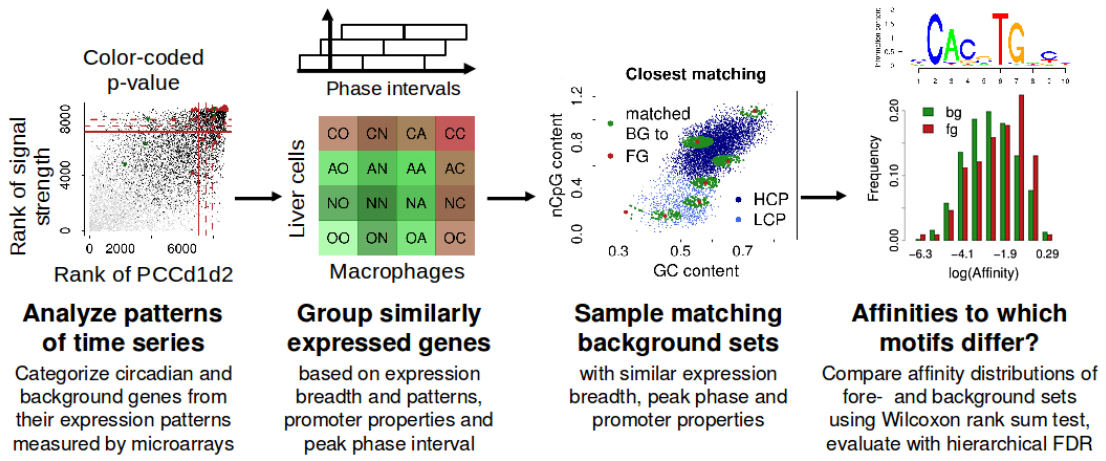


Figure 5.1: **Workflow of analysis.**

As transcription factors bind to highly specific binding sites within the promoter region of target genes to regulate their expression, it is now established that binding sites for common regulating transcription factors are overrepresented in co-regulated genes (Meng et al. [2010]). Numerous studies used the approach of overrepresented binding sites to predict transcription factors involved in the regulation of gene sets (Chang et al. [2006], Roider et al. [2009b]). Here, this method shall be applied to circadian genes, which share the property of a 24-hour period in their expression pattern, while differing among each other in phasing and tissue-specificity of their expression, amplitude and expression level. Thereby experiences of previous analyses (Bozek et al. [2007], Bozek et al. [2009], Yan et al. [2008]) will be further developed.

To find out which motifs are responsible for the tissue-specificity of circadian gene expression or expression timing modulation by overrepresentation analysis, proper background selection is an important matter. To base the choice on the biological question, let's separate the two aspects of the question: tissue-specificity and timing. The focus of the first aspect is to distinguish tissue-specific influences on circadian gene expression from those active in both cell types. In this regard non-expressed genes determined in

each dataset serve as background. Considering timing, we ask for transcription factors with time-dependent abundance, binding or transactivation activity. The answer to this question is rather reflected in the comparison of circadian to non-circadianly expressed genes. However, fore- and background gene sets of mouse macrophages and liver cells overlap and differ to different amounts. Hence, to dissect common and tissue-specific influences subgrouping of genes is necessary. In detail, the idea of subset analyses is encouraged by the following reasons:

(1) To detect transcription factor binding sites which are responsible for the genes' rhythmic expression, promoters of circadian genes shall be compared to those of non-circadian genes. However, a gene detected as non-circadian in one cell type is not determined to be non-oscillating in all tissues. This implies the necessity of a careful background choice, in which subsets of circadian genes are associated to subsets of non-circadian genes in the cell type of interest.

(2) Clearly, circadian genes cannot all be regulated in the same way, because their peak timing is gene-specific. One gene might peak even at different times in two separate cell types. Assuming that the circadian activity of certain transcription factors will be reflected in time-dependent binding site overrepresentation, genes are also grouped by their expression timing into phase interval groups.

(3) Connatural to the search for tissue-specific clock-regulators is the search of tissue-specific transcription factors. As sequence properties of promoters have been associated with expression level and tissue-specificity of gene expression, the promoters' contents of guanine and cytosine base pairs (GC) as well as cytosine-guanine-dinucleotides (CpG) should be considered. To do so, promoters of all genes are classified according to their properties in CpG depleted (LCP) and CpG rich (HCP) promoters. As suggested in the literature (Roider et al. [2009a], Bozek et al. [2009]), these classes are analyzed separately.

In sum, to answer biological questions on the tissue- and phase-specific regulation of circadian genes, they are grouped by tissue-specificity, peak phase interval and promoter property to form foreground sets in subsequent promoter analyses. The next sections give details on the group selection criteria: (1) Are the genes circadian or not in a cell type of interest (section 5.1)? (2) How do gene categorizations compare among two separate cell types (section 5.2)? (3) Within which phase interval peak the gene's expression (section 5.3)? (4) Which gene promoters have low or high CpG content (section 5.5)?

The other sections focus on overrepresentation analysis. Section 5.4 defines the data to be compared between fore- and background sets based on Wilcoxon statistics. The test method and the sampling of suitable background genes to a certain foreground set are described in section 5.6. Due to the high number of subgroup comparisons, a multiple testing procedure is needed for motif prediction. It estimates the proportion of false discoveries among significant results in hypothesis families and will be described in section 5.7.

5.1 Time series analysis filters expression patterns

In order to do binding site statistics on promoters of circadian, non-circadian and non-expressed genes all detected genes have to be arranged into these categories based on recorded time series data. Regarding the selection of background genes this idea is new since previous studies (Bozek et al. [2009]) compared promoters of selected circadian genes with the promoters of all other genes in the database. The disadvantage of this approach is a background promoter set containing promoters of genes with unknown expression levels and patterns in other tissues. In contrast, I use the microarray data to classify fore- and background genes into tissue-specific subgroups within two cell types. Therefore the expression pattern analysis here focuses on the following three questions: (1) Which genes are expressed in the cell types of interest (liver, macrophages)? (2) Which genes show circadian expression profiles? (3) Which genes are expressed in a non-circadian manner?

There is no exact definition of what circadian or non-circadian expression profiles look like. Usually circadian expression profiles are compared to sine waves with a period of 24 hours, but other appearances are also possible, e.g. daily peaks or a saw tooth like pattern (Yang and Su [2010]). A non-circadian pattern is expected to present itself as a series of white-noise-like measurements of expression level, but it could also peak at irregular times due to the influence of regulatory instances apart from the clock.

Based on these ideas several criteria were used to characterize the detected expression patterns and evalulate their biological relevance (see chapter 6): (1) expression level, (2) p-value for significance of rhythmicity, (3) daily pattern similarity and (4) signal strength. In combination, these criteria are employed to assign genes in each cell type into four categories as illustrated in figure 5.2: circadian genes (C), genes ambiguous with regard to circadian or non-circadian character (A), non-circadian genes (N) and non-expressed genes with expression levels below a heuristic expression threshold (O).

5.2 Considering tissue-specificity by category combination subgroups

The comparison of pattern categories assigned to the genes in macrophages and liver cells reveals a large group of detectable genes non-expressed in both cell types (OO) as well as common (CC) and tissue-specific (TS) circadian genes. Among the latter ones are genes with different tissue-specific classifications: (1) tissue-specifically regulated circadian genes (CN/NC), which are circadian in the cell type of interest while non-circadianly expressed in the other one, and (2) tissue-specifically expressed circadian genes (CO/OC), which are circadian in the cell type of interest while non-expressed in the other one. Figure 5.2 shows all possible tissue-specific category combination subsets as overlaps of the pattern categories observed in single cell types.

The existence of these subgroups poses several questions: Is it possible to discriminate between binding sites necessary for expression and those necessary for the circadian oscillation of expression? Do transcription factor binding sites differ between genes

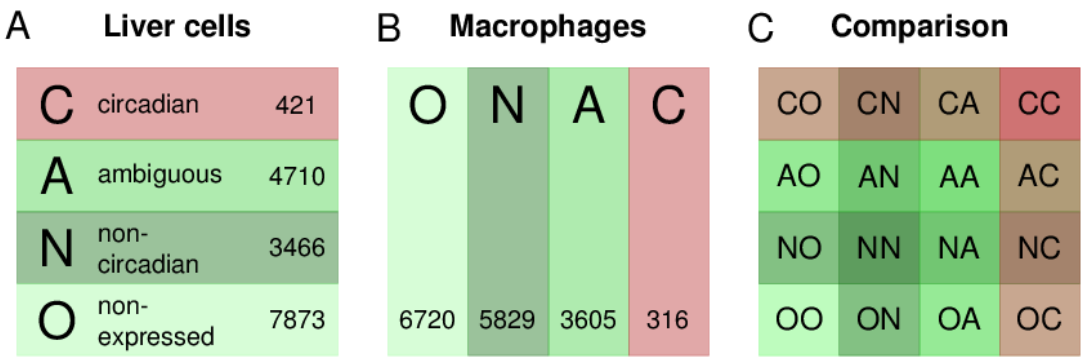


Figure 5.2: **Expression patterns of genes are classified into four categories.** Three of them include expressed genes with circadian (C, *red*), ambiguous (A, *lightgreen*) and non-circadian (N, *darkgreen*) timeseries, the last category contains non-expressed (O, *palegreen*) genes. Shown are the sizes of these categories in liver cells (**A**) and macrophages (**B**). Overlapping these tissue-specific categories yields the subsets relevant for later overrepresentation analysis (**C**). Numbers refer to subsets of genes detectable by both microarray platforms.

which are circadian in several tissues or in only one tissue? It is hypothesized that the regulatory circuits of such tissue-specific circadian genes contain elements common to both cell types as well as cell type-specific elements (Masri and Sassone-Corsi [2010]). However, it is not clear, whether tissue-specific rhythmic regulation in one cell type adds to common non-rhythmic transcription activation or whether common rhythmic regulators are in one cell type contradicted by tissue-specific factors in anti-phase.

With view on tissue-specificity, we want to know which transcription factors contribute to the regulation of genes that show different patterns in the two cell types compared. To answer this question, we use the category combination groups outlined above and change the point of view to the transcription factors' binding sites: Wich set of binding sites can be used to discriminate between circadianly expressed and background genes in cell type one, while in the other cell type both gene sets fall into the same category? Following this question, category comparison setpairs are defined: Promoters of circadian and background genes of the first cell type, which are all in the same category of the second cell type (C, A, N *or* O), are compared with respect to the binding affinities of transcription factors (foreground to background examples: CC to NC or OC, CA to NA or OA, CN to NN or ON, CO to NO or OO). Additionally, tissue-specifically oscillating genes (CA, CN, CO) are pooled (together abbreviated with TS) and compared to background genes of the same cell type that do not occur as circadian in the other cell type. Furthermore, the whole set of circadian genes in one cell type (WH) is compared to all background genes of the same cell type. Alltogether, six possible foreground gene sets in one cell type are defined and will be referred to in the following as category comparison sets (e.g. in liver cells: CC, CA, CN, CO, TS=CA $\cap$ CN $\cap$ CO, WH=TS $\cap$ CC). Conclusions based on the sizes of these subgroups are discussed in section 6.8.

5.3 Considering timing specificity by phase groups

As known from Storch et al. [2002], circadian gene expression accumulates in different phases of the day, depending on the tissue under observation. This is also observed when comparing the distribution of peak phases in gene expression among mouse macrophages and liver (figure 5.3B). To calculate at which phase circadian genes in mouse liver and macrophages have their highest expression level, parameters from the fit equation 6.1 were used: $\omega t = \arctan(A/B)$. The observed difference of peak phase distributions must be caused by tissue-specific transcription factors, because common circadian genes show high similarity of their peak phases (observed here as shown in figure 5.3A as well as by Yan et al. [2008]) indicating similar expression regulation in both cell types.

To capture binding motifs of transcription factors directing phase-specific gene transcription, all genes are sorted into phase groups based on their expression profiles. This is done independently of the time series analysis to ensure that the phase group membership of non-circadian genes is random, while circadian genes are part of only those phase groups around their expression peak. The number of phase groups equals the number of measurements per day. A phase group contains all genes whose expression level at this time of day is higher than their means at one or both days of measurement (figure 5.3C). Within each phase group promoters of circadian genes are compared to those of background genes. Phase groups are considered within each category comparison.

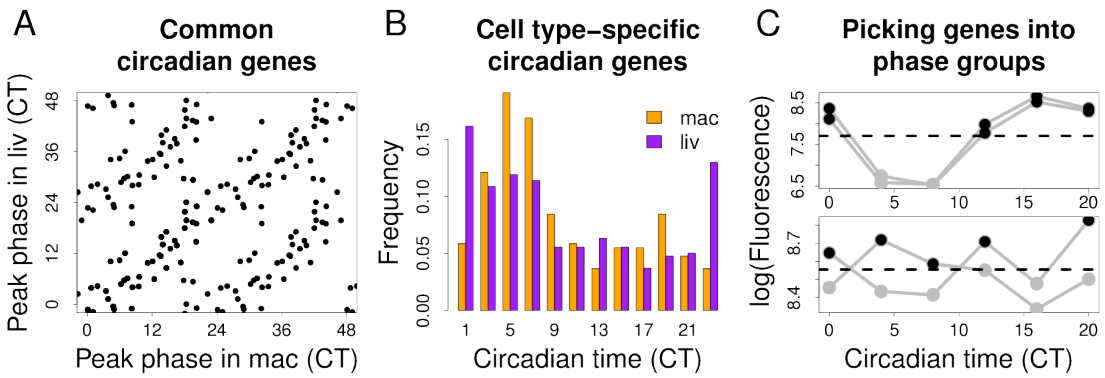


Figure 5.3: **Tissue-specific circadian genes determine peak phase distribution.** **A** Common circadian genes (CC) have similar phase in both cell types, as visualized by the doubleplot of their peak phases in mouse macrophages and liver cells. **B** Peak phase distributions of tissue-specific circadian genes (TS) in mouse macrophages and liver cells differ. **C** Two examples illustrate how genes are chosen into phase groups based on their expression profile (*grey*). Any expression level larger than the mean of the profile (*dashed line*) during the two days of measurement determines phase group membership (*black dots*).

5.4 Predicting transcription factor binding to promoters

Having separated circadian, non-circadian and non-expressed gene groups, the next step is to analyze their potential to be bound by regulating transcription factors. The binding affinity of transcription factors to promoter regions of genes is used as a measure for this potential. For its calculation the matrices of TRANSFAC 12.1 database - representations of the preferred binding sites for most known transcription factors - are used. All promoter sequences were downloaded from the ENSEMBL 58 sequence database.

The reported extent of the promoter sequence that is to be covered in the search for binding sites varies in the literature. As the binding sites of many transcription factors cluster rather closely to and symmetrically distributed around the transcription start site (Consortium [2007]), it is reasonable to restrict the sequence of interest to 1000 base pairs surrounding the transcriptional start site (TSS) by 500 base pairs each up- and downstream. ENSEMBL annotates only the most upstream TSS to genes which have several TSSs. As long as other TSSs are in proximity, they are covered by the defined sequence range for promoter analysis (see section 6.1).

Transcription factor affinities to each promoter were calculated using the program ANNOTATE (previously called TRAP - TRanscription factor Affinity Prediction, Roider et al. [2007]). The affinity value for a particular binding motif to a promoter sequence represents a weighted sum of motif matching scores to all possible sites in the promoter based on a biophysical model. In contrast to the digital hit-counting (Rahmann et al. [2003]) this method does not specify a number of binding locations. It is rather an estimate for the number of binding transcription factors preferring that binding site (section 3.2).

ANNOTATE is successively provided with (1) a tabular description of each documented motif, (2) the promoter sequence of each gene of interest together with (3) its GC content. The last information is important for adjusting the GC content of the randomly generated background when calculating positional weight matrices. The implementation of ANNOTATE into the needs of this study results in a large table listing affinities for all motif matrices to the promoter of each gene detected by any of the two types of microarrays.

5.5 Transcription factors distinguish between promoters with high and low CpG content

As known from recent publications, genes are differently regulated depending on the GC and nCpG content in their promoters (section 2.1.2). The nucleosome-free region in promoters with high GC and nCpG content (HCP) ensures easy access for transcription factors to binding sites and promotes broad expression across several tissues. In contrast, promoters with low GC and nCpG-content (LCP) more often need the tissue-specific removal of a nucleosome because it blocks the transcriptional start site.

To distinguish between these two groups of promoters and avoid additional complexity, I applied a simple method based on the overall content of single guanine and cytosine

nucleotides ($GC = (p(G) + p(C))/n$) as well as the normalized CpG dinucleotide frequency ($nCpG$) of each promoter sequence of length n. The latter is calculated as the ratio of observed CpG dinucleotide frequency to the one expected from the number of present guanines and cytosines by random combination:

$$nCpG = \frac{p(CpG)}{p(C)*p(G)} \quad \text{with} \quad p(CpG) = N(CpG)/(n-1) \tag{5.1}$$

The sum of both promoter properties is used as a promoter property index (PPI) to distinguish between LCP and HCP genes (figure 5.4A):

$$\begin{array}{ll} \text{LCP genes:} & PPI = GC + nCpG <= 1 \\ \text{HCP genes:} & PPI = GC + nCpG > 1 \end{array} \tag{5.2}$$

The abbreviation ACP will cover all genes, that is the union of LCP and HCP genes.

Interestingly, the ENSEMBL database contains much more LCP than HCP genes, although in most tissues more HCP than LCP genes are expressed. Since the ratio of expressed LCP to HCP genes is tissue-specific (Roider et al. [2009a]), I assume that tissue development and cell type maturation play important roles in determining this ratio. Transcription factors driving gene expression in a certain cell type must be able to more easily switch on HCP gene transcription while they should be more selective for LCP genes. Hence, transcription factors need to distinguish between promoters with low or high nCpG content. To see whether they are able to do so, the affinity distributions for each motif are compared between LCP and HCP genes by one-sided Wilcoxon rank sum test followed by p-value correction according to Benjamini and Hochberg [1995] due to multiple testing. It results for almost all transcription factor binding sites in significant differences (FDR<0.05). As an example, panels C and D in figure 5.4 show how small changes in a positional frequency matrix for the transcription factor USF leads to (1) a large change in affinity levels and (2) different affinity distributions for LCP and HCP genes. The upstream stimulating factor USF competes with CLOCK/BMAL1 for binding at Ebox motifs and is involved in lipid and carbohydrate metabolism as well as other cellular processes (Shimomura et al. [2013]).

Based on their preferences for a certain promoter property motifs are divided in two groups named “facilitator motifs” and “specifier motifs”. These names are chosen following Ravasi et al. [2010], who defined transcription factors with widespread expression as facilitators and those with high tissue-specificity as specifiers. Here, in contrast, the names refer to the binding motifs of the transcription factors, but the result is similar. Transcription factors binding facilitator motifs are found to have higher affinities to promoters of broader expressed HCP genes, while more tissue-specific low CpG promoters are enriched among the promoters with highest affinities to transcription factors binding specifier motifs. This is reflected by the sign of the correlation between the logarithmic affinity and the promoter property index (sum of GC and nCpG content, see figure 5.4B).

As discussed in sections 7.2 and 7.3, results of TF affinity comparisons between circadian and other genes may be biased due to different GC and nCpG content distributions among the gene sets compared. Hence, all comparisons must include a control to exclude

such a confounder. This is done by background matching as described next.

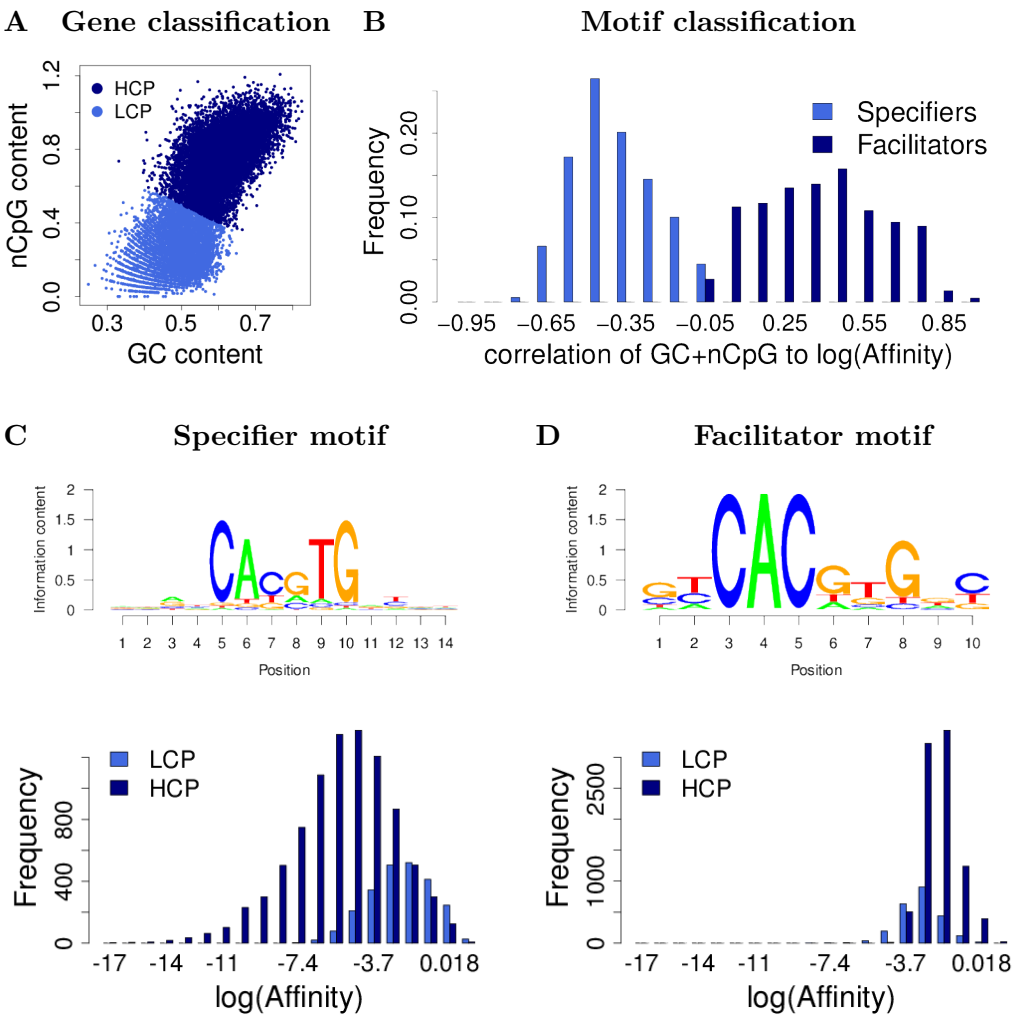


Figure 5.4: **Information content variations lead to affinity differences among promoter classes**

A All detectable genes are classified with respect to GC and nCpG content to LCP and HCP groups. **B** All motifs are classified to specifier or facilitator motifs. Their promoter affinities correlate negatively or positively to the promoter property index (PPI), respectively. **C, D** Sequence logos of two exemplary motifs with the same consensus sequence (canonical Ebox) but different information content at particular nucleotide positions are shown: USF_02 as a specifier and USF_Q6 as a facilitator motif. Small changes in the information content at certain positions affect the TF affinity to LCP and HCP genes: Binding transcription factors have in HCP genes much higher affinity to facilitator motifs than to specifier motifs, while they prefer the specifier motif over the facilitator motif in LCP genes.

5.6 Overrepresentation analysis requires careful background choice

Are there differences in the binding potential of transcription factors to genes of two groups? In particular, do motifs exist, to which foreground gene promoters have a higher affinity than background gene promoters? Overrepresentation analysis is a method to answer such questions. Affinity distributions of promoters in a circadian foreground set and a 500 times larger background set containing randomly sampled non-circadian or non-expressed genes are compared using the non-parametric one-sided Wilcoxon rank sum test (see section 4.4.3). An exemplary foreground-background comparison is shown in figure 5.5.

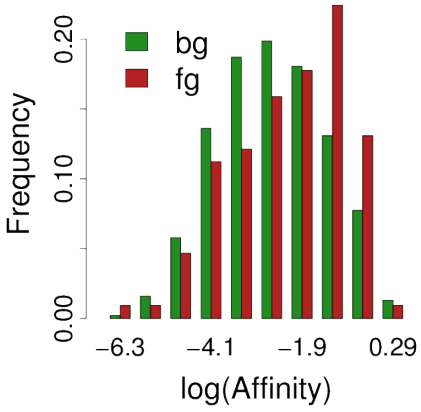


Figure 5.5: **Example Wilcoxon test.** Tissue-specific circadian liver genes (TS) with high normalized CpG content are at CT14 enriched with Eboxes of the general kind compared to matching non-circadian HCP genes. Affinity distributions of transcription factors binding to the motif EBOX_Q6_01 (consensus sequence "CANNTG") in foreground and background genes are shown in *black* and *grey*, respectively. Wilcoxon test statistic indicates a significant difference ($p<0.001$).

To find binding sites responsible for the tissue-specific cicadian regulation of genes in separate cell types, comparable promoter sets must be selected. The measurable circadian oscillation of a gene's expression in one cell type indicates that its sequence possesses binding sites enabling rhythmicity. However, the same gene may be non-expressed, non-circadian, circadian or expressed with ambigous pattern in the other cell type. Accordingly, a non-circadian or non-expressed gene in one cell type may belong to one of the four mentioned categories in the other cell type. Asking for cell type-specific regulators of rhythmicity implies to compare subsets of circadian and background genes in the cell type of interest which are assigned to one and the same category in the other cell type.

When calculating the affinity of a transcription factor to a promoter region, the effect of present transcription factor binding sites intermingles with the effect of promoter properties in the final affinity value (section 5.5). As this study aims to find overrepresented transcription factor binding sites based on a comparison of affinity distributions between circadian and non-circadian or non-expressed gene sets, it is essential that the sequence parameters GC and nCpG content are distributed as similarly as possible within fore- and background sets to exclude false discoveries based on such distribution differences. To control the GC and nCpG content distributions in background sets, a matching procedure is used. The following subsections explain this procedure.

5.6.1 Why the expression level of background genes matters

Previous analyses compared promoters of circadian genes to all other promoters in the database (Yan et al. [2008], Bozek et al. [2009]). They left aside knowledge about tissue-specific expression levels and patterns in background genes. To overcome this problem, I selected foreground as well as two types of background gene pools (non-expressed and non-circadian) based on microarray data. Thereby I disentangle the influences of transcriptionn factors on tissue-specificity and timing. This strategy will help to identify possible tissue-specific binding partners for clock regulators assumed by Masri and Sassone-Corsi [2010].

The database contains lots of non-expressed genes whose expression may oscillate in a different tissue not under observation in a study of interest. Based on their background choice it is plausible that Bozek et al. [2009] found a higher content of guanine and cytosine bases (GC) in circadian genes, because the latter are mainly expressed. As illustrated in figure 5.6, expressed genes generally have higher GC and nCpG contents than non-expressed genes. This underlines the importance of proper gene set selection for promoter analysis. How overrepresentation results change depending on the used background pool will be discussed in section 7.3.4.

Comparing circadian and non-circadian among expressed genes, GC and nCpG contents are distributed more similarly (figure 5.6). However, many phase-specific subsets of circadian genes with tissue-specific rhythmicity of expression profiles differ significantly from their proper background sets with respect to at least one of the promoter properties (section 5.6.3). To treat gene sets in all tests the same, background sets are sampled from the chosen background pool to match GC and nCpG distributions of foreground sets as described in section 5.6.2.

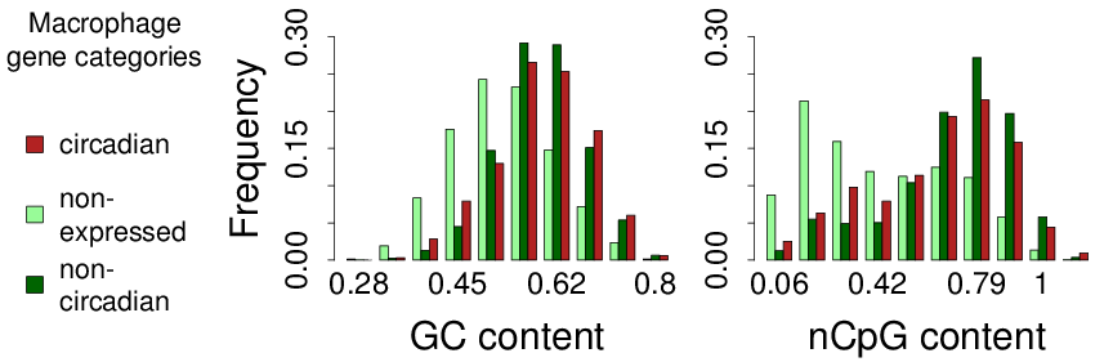


Figure 5.6: **Promoter properties vary with expression level.**
Distribution differences of GC and nCpG content in gene sets are much stronger when comparing circadian to non-expressed genes than to non-circadian genes. Data from macrophage circadian, non-circadian and non-expressed categories are shown as representative example.

5.6.2 The procedure of background matching

Due to general differences in the distribution of promoter properties in foreground and background gene sets used for comparison, overrepresentation analysis may detect motifs falsely. To exclude such results within the limits of control, it is essential that GC and nCpG contents are distributed as similarly as possible between the compared sets. Therefore, I fine-tuned a matching procedure introduced by Bozek et al. [2009]. Since a matching of nCpG content also impacts the distribution of GC content, matching is done for both criteria simultaneously.

To control the distribution of GC and nCpG content in randomly sampled background sets, for each foreground gene (fg) matching background genes (bg) are sampled with replacement from the most similar genes with regard to GC and nCpG contents. For genes with less dense neighborhood, at least ten closest matching genes served as choosing pool, otherwise all genes with a distance of less than 0.05, with the distance d being defined by:

$$d = \sqrt{(GC_{\mathrm{fg}} - GC_{\mathrm{bg}})^2 + (nCpG_{\mathrm{fg}} - nCpG_{\mathrm{bg}})^2} \tag{5.3}$$

To test whether this matching procedure indeed samples background sets with GC- and nCpG-distributions resembling the ones of the foreground set, the Kolmogorov-Smirnov-test is applied (section 4.4.5). For both promoter properties random gene sampling without matching leads to significant differences of the fore- and background distributions, while background sets sampled according to the matching procedure can be assumed to be taken from the same population as the foreground set (figure 5.7).

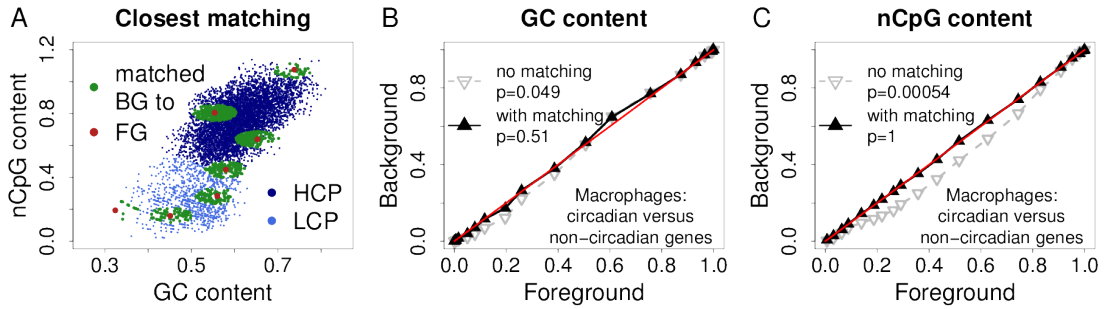


Figure 5.7: **Background matching increases similarity of promoter property distributions between compared promoter sets.**
A For exemplary LCP and HCP foreground genes the closest matching non-circadian genes with similar sequence properties (distance $d \leq 0.05$, at least 10 genes) are marked. From each foreground gene's matching set 500 genes are chosen randomly with resampling into the background set. **B, C** Resulting background sets after sampling non-circadian genes without and with matching to macrophage circadian genes are compared to the foreground set with respect to the promoter properties GC and nCpG content. The QQ-plots illustrate that the matching procedure eliminates largely previously significant differences. P-values report the results of Kolmogorov-Smirnov-tests.

5.6.3 Necessity of matching

How different are circadian foreground and their background genes with respect to their promoter properties? To answer this question, GC and nCpG distributions of foreground subsets in all category combination and phase groups are compared to the ones in their respective backgrounds sampled from non-expressed and non-circadian genes. If the two-sided Kolmogorov-Smirnov-test results in significant p-values ($p<0.05$) with respect to at least one of the promoter properties, matching is necessary. Set pairs with this condition are counted. On average, 44% of set pairs need matching, in case of comparing circadian to non-expressed macrophage genes the proportion goes up to 70%. However, to treat all sets the same during the analysis, matching will be applied to all set pairs. Whether this has an impact on the overrepresentation results will be discussed in section 7.2.

5.7 Finding significant results within multiple tests

In sum, circadian genes of mouse macrophages and liver are each divided into six cell type-specific category combination subgroups (ccs: $CC, CI, CN, CO, TS = CI \cup CN \cup CO, WH = CC \cup TS$) within three classes of promoter properties (ppc: *LCP, HCP, ACP*), where the last one is the union of the first two classes. Whithin these 6*3=18 subgroups, genes belong to 6 or 24 overlapping phase groups (phs) in macrophages or liver, respectively. Each of the $6 * 18 = 108$ (in macrophages) or $24 * 18 = 432$ (in liver) phase groups are viewed as a foreground set in turn to calculate one-sided Wilcoxon rank sum statistics on affinities to 611 binding motifs in comparison to their respective background sets, which are chosen from non-circadian or non-expressed genes in separate analyses. In each test, the Null hypothesis is that the affinity distributions are similar, while the alternative is that affinities of circadian genes are shifted to larger values in comparison to background genes. The Null hypothesis is rejected in favour of the alternative, if the test's p-value falls below the significance threshold α=0.03. Finally, the number of single tests is $18 * 6 * 611 * 2 = 131,976$ in macrophages and $18 * 24 * 611 * 2 = 527,904$ in liver, hence alltogether $659,880$.

Within this number of tests positive results with p-values below the significance threshold α=0.03 are expected to occur by chance, because p-values are distributed uniformly. Assuming that motifs bound by clock-related transcription factors offer higher affinities in circadian genes than in other genes (which are non-expressed or expressed with non-circadian pattern), the Wilcoxon test results are used to identify motifs enriched significantly in certain phase and subset groups. Whether a motif is finally predicted as associated to circadian gene regulation within one and/or the other cell type depends on the number of p-values for this motif that fall below the significance threshold. In one single test motifs may be detected falsely (False Positives, FP, α-error), as many of the other tests per motif result in large p-values. Motifs with too few p-values below the significance threshold may falsely not be associated with circadian regulation (False Negatives, FN, β-error). These possible error types are shown in the contingency table in the left panel of figure 5.8. However, the motif prediction is powerful if possible errors are minimal. In the context here, we are interested in the detection of motifs that might be

Wilcoxon test decision

motif	$p \leq \alpha$	$p > \alpha$	sum
pred.	TP	FP	$D * H^{mtf}$
other	FN	TN	$nD * H^{mtf}$
sum	m_1	m_0	m

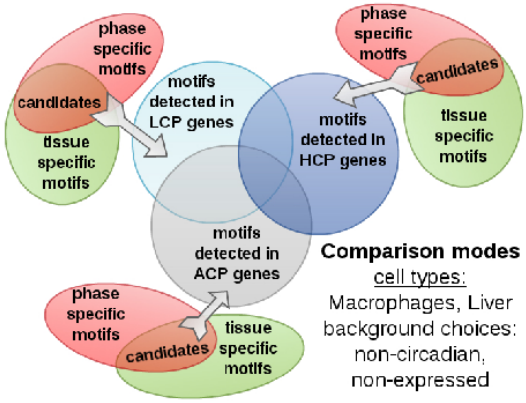


Figure 5.8: **P-values measured in multiple gene set comparisons are used to predict motifs that might affect tissue-specific circadian regulation.** **Left** Contingency table showing the possible truthfulness of test decisions in m tests (TP: true positive, TN: true negative, FP: false positive, FN: false negative). While the number of positive (m_1) and negative (m_0) Wilcoxon test results depending on the significance threshold α is observable, the number of motifs D (each tested in H^{mtf} hypotheses) that can be predicted to affect circadian regulation is estimated based on a hierarchical multiple testing procedure illustrated in the other panel. The FDR of this estimate is $FP/(D * H^{mtf})$. **Right** A hierarchical strategy helps to find motifs involved in tissue-specific circadian gene regulation. To predict candidate motifs involved in tissue-specific timing regulation, promoters of circadian genes determined in two cell types (mouse macrophages and liver cells) are compared to selected background genes depending on the biological question (comparison modes). Asking for tissue-specific circadian expression, promoters of non-expressed genes are used as background, while non-circadian genes serve as background when asking for circadian modulation of expression patterns. Each analysis reveals in LCP and HCP genes separately as well as for all genes together (ACP) motifs which are enriched in certain phase and category combination subgroups of circadian genes below the significance level α. The overlaps of all comparison modes will be discussed in section 7.4.

involved in tissue-specific timing regulation. In this context, a false detection is not considered as a serious mistake, as long as many other hypotheses are true. Hence, among all the possible motif discoveries, the number of false discoveries should be controlled (α-error). An appropriate procedure is in this case the False Discovery Rate (FDR). It can be described as the "average truthfulness of the selected hypotheses" (Rosenblatt [2013]) and is calculated as the proportion of false discoveries in all discoveries.

However, using the procedure for p-value adjustment described by Benjamini and Hochberg [1995] (BH, see section 4.4.2) on all p-values pooled together, the number of corrected p-values smaller than the cutoff α=0.03 is zero. This is, because due to the previous grouping of genes by tissue- and phase-specificity hypotheses are not exchangable.

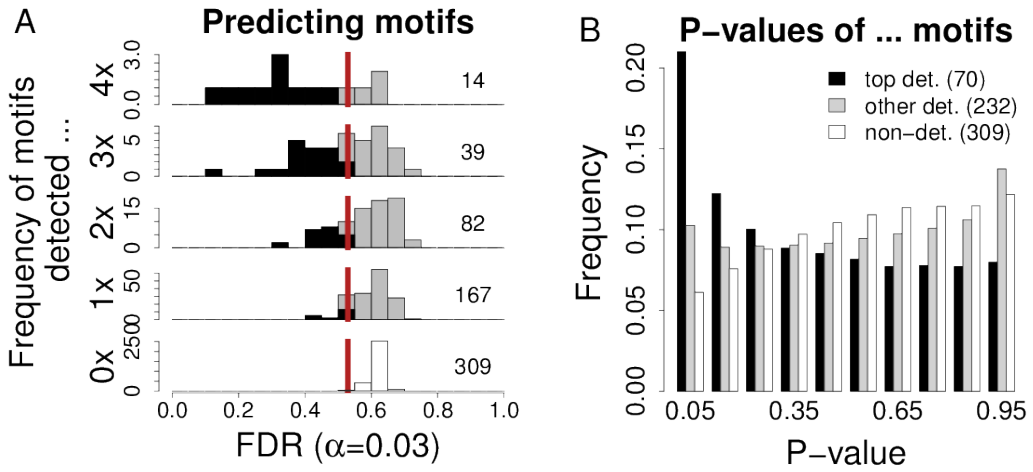


Figure 5.9: **Distributions of FDRs and p-values of predicted motifs.** **A** Motif FDRs are inversely proportional to the number of comparison modes the motifs are detected in. Each histogram refers to the FDRs of motifs detected in four to zero comparison modes. The number to the right of each histogram indicates how many motifs are detected in as many comparison modes. The FDR cutoff ϵ=0.53 is marked as *red line* (details in section 7.1). **B** Distributions of all Wilcoxon p-values measured in macrophages and liver for TP=70 predicted, FN=232 and all other m_0=309 motifs are shown. The color code is the same in both panels.

For example, gene sets sourced from two different category comparison groups and phases timed 12 hours apart (e.g. CT0 from CO and CT12 from CC) are not expected to be regulated by similar transcription factors. According to Efron [2008], a pooled analysis is not appropriate here.

Instead, all hypotheses for each motif are grouped by promoter class and within those by families according to the two questions: (1) May the motif be important in a certain phase group (phs)? (2) Does the motif occur especially often in a certain category combination subgroup (ccs) of one promoter class? Candidate motifs give affirming answers to both questions in at least one promoter class. This strategy, a hierarchical multiple testing procedure, is applied in four comparison modes using circadian genes of mouse macrophages and liver cells as foreground and respective non-circadian and non-expressed genes as background in turn. It is illustrated in the right panel of figure 5.8 and described in more detail in section 7.1.

This way, m_1=302 motifs are detected in at least one of the four comparison modes for cell type specific regulation of circadian genes in macrophages or liver. Most of them are found in either LCP or HCP gene subsets; the majority of motifs found in gene sets containing both promoter groups are also found in the subclasses (figure 7.9). From this point of view the division of gene sets into promoter property classes was advantageous.

The false discovery rates of detected motifs are inversely proportional to the number of comparison modes the motifs are detected in. The FDRs of TP=70 motifs fall below the

threshold ϵ=0.53 where it is controlled at significance level α=0.03 (black in figure 5.9A). Among all Wilcoxon hypothesis tests carried out in this study these motifs achieved the most significant p-values (figure 5.9B). Binding transcription factors to these motifs are listed in table 7.1, while those binding to the other FN=232 motifs detected more tissue-specifically are listed in table 7.2.

Motifs enriched in circadian genes of macrophages and liver independent of background choice include the canonical Ebox binding site "CACGTG", to which the oscillating transciption factor CLOCK:BMAL1 binds. Many related general Ebox motifs with consensus "CANNTG" are found more tissue-specifically, suggesting that tissue-specific binding partners or competitors determine tissue-specific circadian regulation. A similar observation was already reported on the NF-kappaB binding site which is able to change the binding factor's dependence on a specific coactivator by a single nucleotide change (Natoli [2004]). Detailed results of the promoter analysis are reported in chapter 7.

6 Characterization of circadian and non-circadian gene expression

The following analysis aims to classify genes as non-expressed or as expressed in a circadian or non-circadian manner. It is based on gene expression profiles that were measured by microarray time series (Hughes et al. [2009], Keller et al. [2009]). In short, at several timepoints nuclear mRNA is extracted from collected cells, amplified via polymerase chain reaction, labelled with a fluorescent dye and hybridized to DNA oligomers that are spotted on and covalently bound to a small glass plate, the microarray. After washing, the spots where hybridization took place can be detected by fluorescence microscopy. To use several microarrays in a common time series analysis, they are normalized. One datapoint is calculated from a probeset designed of several oligomers containing sequences complementary to a target DNA and sequences with mismatches to estimate nonspecific binding. The resulting tables of fluroescence measurements for each probeset at all timepoints are provided by Dr. Kuban for macrophages and downloaded from the Gene Expression Omnibus for liver cells.

6.1 Probeset mapping to genes

Using the annotations listed in the Ensembl 58 database probesets of Affymetrix microarrays were mapped to Ensembl gene entries. Probesets which are annotated to several different genes as well as those which presumably crosshybridize according to the probeset description given by Affymetrix were discarded. If a group of probesets concertedly detected the same gene the one showing the best circadian expression pattern was chosen to represent the particular gene's expression pattern. In order to do so, from several characterization criteria (sections 6.2 to 6.5) the best was chosen hierachically: (1) lowest p-value, (2) highest sum of ranks of signal strength and day1-to-day2 correlation, (3) highest median expression level. In addition to these criteria also promoters with nonamed bases ("N") within their sequences were excluded from further analysis.

Numbers of probesets, detected and expressed genes The microarrays used to detect expression profiles in macrophages (mac) and liver cells (liv) contained 35557 and 45101 probesets, respectively. Sorting out cross-hybridizing probesets and annotating probesets to genes in the database Ensembl 58 resulted in 23049 (mac) and 29000 (liv) non-ambiguous probesets. Those probesets detected alltogether 22304 (mac) and 17463 (liv) genes. For genes which were detected by several probesets (460 in macrophages, 520 in liver), one probeset was chosen to represent the gene's expression profile according to the description above. Thirteen (mac) and sixteen (liv) promoters were excluded due to

nonamed nucleotides within their sequence. Finally, 22291 (mac) ad 17447 (liv) genes of ENSEMBEL 58 were overlapped to find 16470 genes that could be detected by both array platforms. Of those, 10863 were expressed in macrophages and 8850 in liver cells. Expression of 7152 genes was detected in both tissues according to the median cutoffs mentioned in section 6.2.

Transcription start sites Sequence downloads refer to the transcription start site (TSS) annotation from Ensembl 58. For each gene the most upstream TSS annotated was used.

Many genes have alternative promoters, but it is not clear which one was responsible for the observed expression profile. To find out, how many of these alternative promoters are captured in this analysis, all TSSs annotated in ENSEMBL 57 were downloaded and their distances to the most upstream TSS measured. They reached from 1 to 2,184,341 bp with small distances being much more frequent. The distribution of these distances shows, that about 35% of all alternative TSSs lie inside the range of 500 base pairs downstream of the one that was chosen for this analysis.

6.2 Focussing on overcritically expressed genes reduces noise

When measuring expression profiles of several tens of thousands of genes, one captures expressed as well as non-expressed genes, since only about 10,000 genes are active in one cell type. A certain number of detectors on a microarray will also crosshybridize to several mRNA molecules. The less abundant a specific hybridizing mRNA, the higher is the likelihood of crosshybridization. To exclude non-expressed genes from further analysis, I used the median expression level of the log-transformed data.

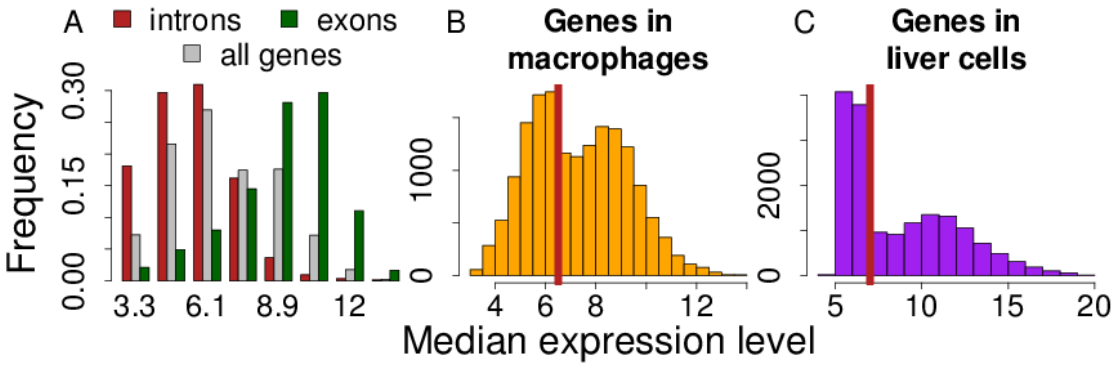


Figure 6.1: **Levels of gene expression are bimodally distributed.**
A Probesets hybridizing to exonic and intronic DNA regions as labeled in the macrophage dataset justify the association of the two modes to expressed and non-expressed profiles. Histograms of median gene expression levels are shown for mouse macrophages and liver cells in panels **B** and **C**, respectively. Cutoffs used in the following analysis are indicated in *red*.

The histogram of median expression levels shows a bimodal distribution (figure 6.1). This is typical for data of microarrays which were normalized using the GeneChip Robust Multiarray Averaging (GC-RMA) procedure, because this method considers each probe's sequence and GC-content for background correction of non-specific hybridization measurement. Comparing the distributions of median expression levels measured by control probesets targeting exonic (expressed) and intronic (non-expressed) DNA regions shows that these are strongly shifted apart from each other according to their expression fate. To exclude rather noisy expression patterns I abandon all time series with a median lower than a certain detection level (expression cutoff). It was chosen to be 6.5 in mouse macrophages and 7 in murine liver cells.

In the later sections it will be shown that expressed genes have a higher enrichment of circadian profiles than non-expressed genes. It seems that the expression profiles above and below the expression cutoff rather result from active biological control and noise respectively.

6.3 Comparison of harmonic to constant fits identifies sinusoidal patterns

The intuitive expectation on a circadian expression profile is a sine wave. It oscillates symmetrically around its mean level. Many circadian genes show a peak-like expression pattern on the fluorescence scale. This appears more similar to a sine wave when transformed with the logarithm to the base of two. Therefore, logtransformed data were used for the following analyses.

A sinusoidal wave function $wave(t)$ is fitted to the logtransformed time series data $x(t)$ with the mean expression level $\langle x \rangle$ (dashed blue line in figure 6.2) according to the equation:

$$wave(t) = \langle x \rangle + A \sin(\omega t) + B \cos(\omega t) \tag{6.1}$$

This function has three parameters (p_{wave}). When fitted to a time series consisting of n datapoints, the model has $n - 3$ degrees of freedom.

In contrast, the expression profile of a gene that is not regulated by the circadian clock is generally associated with constant expression levels or a white-noise-like expression mode. Thus, it would be perfectly fitted by a constant function $mean(t)$ at the mean expression level $\langle x \rangle$ (black line in figure 6.2):

$$mean(t) = \langle x \rangle \tag{6.2}$$

Since this function has only one parameter (p_{mean}), such a regression model to a time series of n datapoints has $n - 1$ degrees of freedom.

Each regression model has a variance describing the average distance of the model to the measured datapoints (illustrated as dotted lines in figure 6.2). It is calculated as the sum of the squared residues (sos_{fit}) from the time series datapoints to the used fit

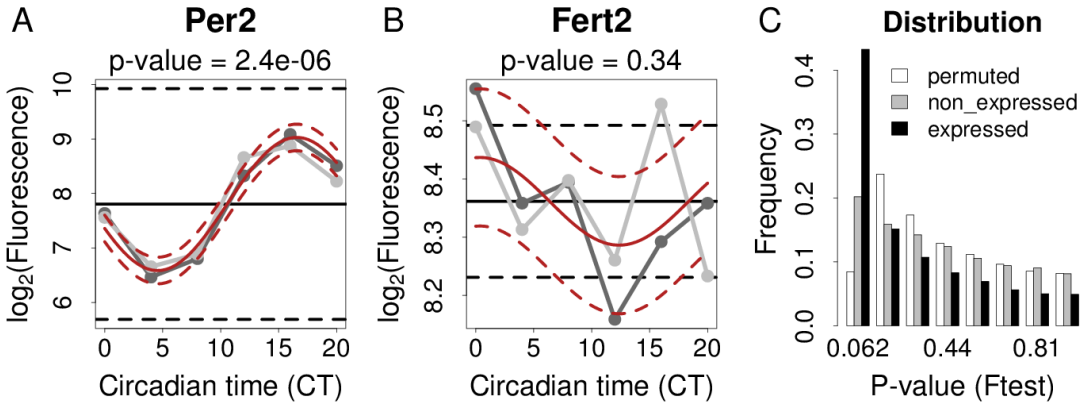


Figure 6.2: **By comparing harmonic to constant fits sinusoidal patterns are found.**
Time series of exemplary circadian (**A**) and noncircadian (**B**) genes are shown from mouse macrophages as overlays of day 1 (*darkgrey*) and day 2 (*lightgrey*). They are fitted with a constant function (*black*) and a sine wave (*red*). Respective variances of these fits as marked with *dashed lines* are compared by Ftest yielding the indicated p-values. **C** The proportion of sinusoidal patterns is higher in profiles of expressed genes compared to non-expressed genes or permuted datasets.

function divided by the models' degrees of freedom ($df_{\text{fit}} = n - p_{\text{fit}}$):

$$var_{\text{fit}} = \frac{sos_{\text{fit}}}{df_{\text{fit}}} = \frac{\sum_{t=1}^{n}(x(t) - fit(t))^2}{n - p_{\text{fit}}} \tag{6.3}$$

The difference between the models is expressed as

$$var_{\text{diff}} = |\frac{sos_{\text{mean}} - sos_{\text{wave}}}{df_{\text{mean}} - df_{\text{wave}}}| \tag{6.4}$$

To decide whether a time series shows a sinusoidal pattern the question is raised whether the *wave* model fits the data better than the *mean* model. However, since the *wave* model uses two more parameters than the *mean* model, it fits any time series better. The *mean* model is nested in the *wave* model, because by choosing particular parameters in the *wave* model the *mean* model can be achieved.

To evaluate which time series is regressed significantly better by a *wave* function F-test statistics is used. It helps to conclude whether the variation in the measured data is due to noise or to timing of gene regulation in a sinusoidal rhythm. Generally, it compares two distributions by calculating their variance ratio, taking into account the degrees of freedom (*df*). In this case, the *mean* model variance var_{mean} is compared to the variance var_{diff} which describes the difference between the two models:

$$F_{\text{wave}} = \frac{var_{\text{diff}}}{var_{\text{mean}}} \tag{6.5}$$

If the resulting F-ratio is larger than the critical value corresponding to the degrees of freedom and the significance level of $\alpha = 0.05$, the null hypothesis of an adequate data fit by the *mean* model is rejected. A p-value is given to each test result.

Examples of circadian (Per2) and non-sinusoidal (Fert2) expression profiles in mouse macrophages and liver are illustrated in figure 6.2 A and B. These profiles are plainly distinguishable by their p-values. Overall, among expressed genes there is a strong enrichment of time series with low p-values as shown for mouse macrophages in the histogram of p-values in figure 6.2C.

However, there are time courses for which the *wave* fit does a significantly better job than the *mean* fit from a mathematical point of view, but the patterns of both days differ too much to convincingly result from biologically rhythmic regulation (see figure 6.3). Additionally, an oscillation of expression levels within a high range seems to be more significant to biological processes than one with a low amplitude (see figure 6.4). Therefore, I introduce two more criteria for the selection of circadian genes and discuss them in the following sections.

6.4 Day-to-day correlation analysis identifies daily repeating patterns

Circadian expression profiles cannot always be described well enough by a sine wave pattern. Some genes peak or dip just during a certain time of day, others show a saw

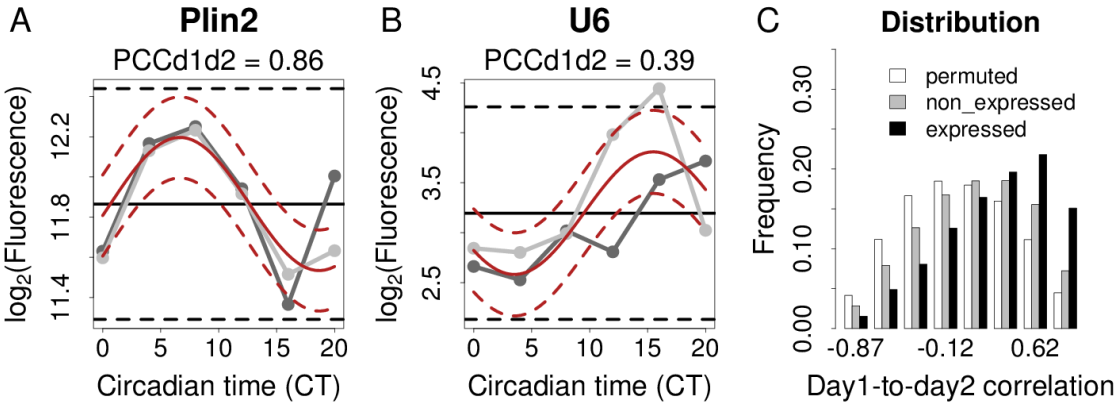


Figure 6.3: **A strong correlation of daily expression patterns is a useful criterium.**
A, B Two exemplary time series from mouse macrophage genes with p-values below 0.02 are shown as overlays of day 1 (*darkgrey*) and day 2 (*lightgrey*). Their constant and sine wave fits are plotted in *black* and *red*, respectively, while the fit variances are dashed. However, their day1-to-day2 correlations differ notably. **C** Day1-to-day2 correlation is distributed to larger values in profiles of expressed genes in comparison to those of non-expressed genes or permuted data.

tooth like expression pattern. What they have in common is a period of around one day with a similar expression pattern each day. To measure the day-to-day correlation of expression patterns, the Pearson Correlation Coefficient is used. Therefore, each logtransformed time series is split in two parts of $n/2$ timepoints (x_1 and x_2), each characterized by a mean $\langle x_i \rangle$ and standard deviation s_{x_i}, which are compared to each other by the Pearson Correlation Coefficient (PCC):

$$PCCd1d2 = \frac{1}{n/2-1} \sum_{t=1}^{n/2} \left(\frac{x_1(t) - \langle x_1 \rangle}{s_{x_1}}\right)\left(\frac{x_2(t) - \langle x_2 \rangle}{s_{x_2}}\right) \tag{6.6}$$

A good correlation of both days ($PCCd1d2 \approx 1$) points to a daily recurring expression pattern, while a missing correlation between time-related measurements results in a $PCCd1d2 \approx 0$. The day1-to-day2 correlation can thus also be seen as a measure for the signal-to-noise ratio, if the p-value of the Ftest is small.

Daily recurring patterns are found more often in the measured data than in permuted time series, as shown in figure 6.3. Expressed genes show more day-to-day similarity than non-expressed genes.

6.5 Signal strength is a useful criterium for biological relevance

A periodical expression profile only seems to be biologically relevant in the context of time dependent gene regulation, if the range of oscillating expression levels is large in comparison to the base expression level of a gene. This may be regulated tissue-specifically. An example of oscillating gene expression profiles with differing signal

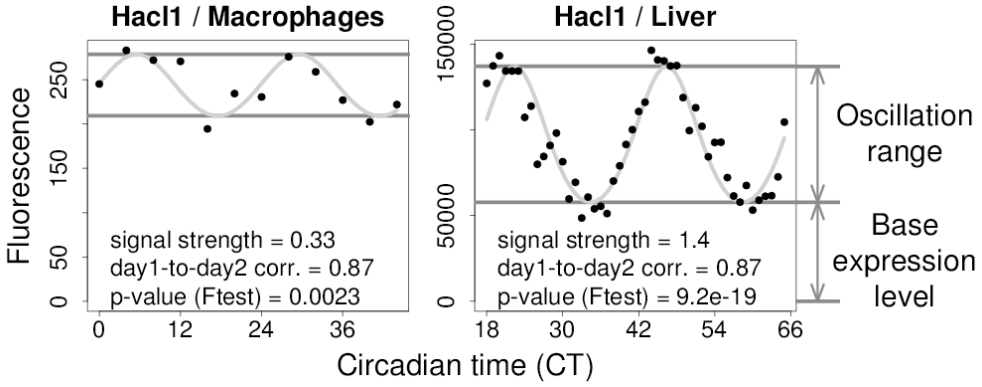


Figure 6.4: **Signal strength is a relevant criterium for circadian genes.**
The gene *Hacl1*, whose product is involved in fatty acid breakdown, is determined as circadian based on its expression profile in mouse liver cells only, because in mouse macrophages the signal strength is too low. Nevertheless, *Hacl1* expression could be regulated by the circadian clock in both cell types. Tissue-specific factors in liver could serve to amplify its oscillation range.

strengths between mouse macrophages and liver cells is shown in figure 6.4.

As the oscillation range is more meaningful on the measured fluorescence scale, expression level peaks and troughs were recalculated using the amplitude $Amp = \sqrt{A^2 + B^2}$ and mean $\langle x \rangle$ given by the harmonic regression (Eq. 6.1) of each logtransformed time series. Taking the peak-to-peak amplitude as oscillation range and relating it to the base expression level (the minimum level of harmonic oscillation) the signal strength is calculated as follows:

$$\text{signal strength} = \frac{\text{peak-to-peak amplitude}}{\text{base expression level}} = \frac{2^{\langle x \rangle + Amp} - 2^{\langle x \rangle - Amp}}{2^{\langle x \rangle - Amp}} \tag{6.7}$$

A larger amplitude brings about a better harmonic regression, this explains the negative correlation between signal strength and p-value. Furthermore, a larger amplitude rhythm can be better distinguished from noise. Therefore, the combination of these criteria may enhance gene selection quality.

6.6 Combining several criteria reduces the false discovery rate

Each of the discussed criteria alone (p-value, day1-to-day2 correlation, signal strength) is not reliable enough to distinguish circadian from non-circadian time series. A profile may be perfectly sinusoidal (indicated by a tiny p-value), but have such a small amplitude that it is not convincing to assume a time-dependent biological functionality for the associated gene. Additionally, genes with daily reoccurring peaking or saw-tooth-like patterns may be missed when choosing circadian genes with a stringent p-value cutoff. On the other hand, a high signal strength may either result from a reliable circadian rhythm recognized by the wave fit or from large noise or extreme outliers. A good day1-to-day2 correlation may result as well from circadian patterns as from patterns with 12 h period or overall trends in the gene expression level. On top of all this, the recognition of a rhythmic pattern in an expression profile is no guarantee for the regulation of this particular gene by the endogenous circadian clock mechnism. The profile may have occurred by chance.

A reliable as well as feasible selection of expressed circadian genes should yield genes with biologically convincing profiles and as little as possible falsely detected genes. Experimentally confirmed circadian genes should be recognized. Furthermore, the use of a standard selection procedure would facilitate the comparison of circadian genes between the two cell types of interest in this study. Unfortunately, it is not known how many genes are really regulated by the circadian clock mechanism. The sizes of finally selected circadian gene sets depend on the chosen cutoffs for each criterium.

Based on their calculation, the three criteria are not completely independent. This is due to the fact that the signal strength is calculated from the amplitude in the harmonic fit to exclude misleading inferences based on outliers. Furthermore a very good harmonic regression implies a good day1-to-day2 correlation and a high amplitude. What is the advantage of using them in combination to choose circadian genes?

While the rankings of the signal strength as well as day1-to-day2 correlation relate

well to the ranking of p-values (Spearman rank correlations in the range of 0.67-0.77), they correlate badly to each other within the set of expressed genes (Spearman rank correlation in the range of 0.2-0.3). Thus, a preselection of profiles for high signal strength and high day1-to-day2 correlation may strengthen the ability of the p-value to discriminate between circadian and non-circadian genes. This was tested by measuring the rate of detected circadian and non-circadian profiles in the experimental as well as in their permuted time series. Comparing the rates of detected profiles in both sets yields the ROC curve in the right panel of figure 6.5. It shows that true detections by the p-value threshold are much more probable when the profiles are preselected by the other two criteria.

The combination of three criteria is a valuable solution for the described selection problem. It offers the possibility to choose less stringent cutoffs for each criterium in order to include profiles with reoccurring patterns that do not match a sine wave very well while ensuring finally a low false discovery rate.

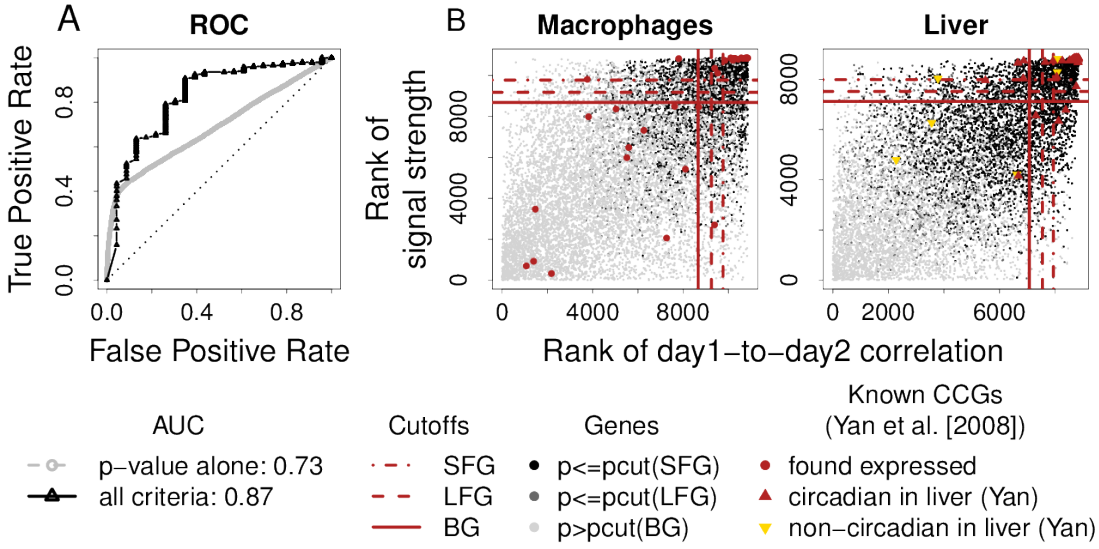


Figure 6.5: **Expression patterns are classified using three criteria.**
A Combined the three criteria p-value, day1-to-day2 correlation and signal strength allow for better discrimination between biologically relevant circadian and non-circadian expression profiles. The ROC curve shows significance and sensitivity of p-value classification alone or together with the other two criteria on expressed genes in macrophages (SFG criteria). **B** The border between circadian and non-circadian profiles is unclear. The dotplots show the indefinite transition from circadian (*corner up right*) to non-circadian (*corner down left*) profiles of expressed genes. For each time series values for the three criteria are plotted as rankings of signal strength and day1-to-day2 correlations on the x- and y-axes, respectively, and color-coded p-values. The displayed cutoffs serve to classify genes into expression pattern categories (see table 6.7). Genes of the list of 41 known circadian genes (Yan et al. [2008]) are highlighted.

6.7 Selected circadian genes include known clock genes

Among expressed genes small and large foreground sets (SFG, LFG) containing circadian genes as well as a background set (BG) with non-circadian genes are defined. As the datasets of mouse macrophages and liver cells encompass differing numbers of time points and ranges of fluorescence measurements, the selected cutoffs differ between them. However, the thresholds for the day1-to-day2 correlation and signal strength are determined in a similar and automated way. Foreground genes are chosen such that their values are ranked among the largest 10% (SFG) and 15% (LFG). All genes with a day1-to-day2 correlation and signal strength lower than the top 20% are regarded as background genes, if their p-values also exceed the background p-value threshold. The exact cutoffs for the small and large foreground sets as well as the background sets are shown in table 6.7. The selection procedure classified all detected genes into four groups: circadian, non-circadian, ambiguous - the three expressed gene groups - and one containing all non-expressed genes.

To evaluate whether the described method recognizes known circadian genes (KCG), the selected lists are compared to a reference list taken from Yan et al. [2008] who measured gene expression profiles in 14 mouse tissues and published 41 genes showing circadian expression in at least eight of them. If those were found expressed in macrophages or liver cells, they are shown as big dots in figure 6.5. In liver, of the 41 KCGs two genes were excluded by the expression cutoff and six were not detected as circadian by Yan et al. [2008]. Two of the latter (*Per3* and *Hnrpdl*) may have been false negatives, as they are detected circadian in my analysis. 70% of the remaining 33 expressed KCGs in liver are detected as circadian in this analysis as well. For macrophages there is no simialar set of known circadian genes published. Nevertheless, the expression profiles of 42% of 40 expressed KCGs are found to oscillate. In sum, the selection result is in good agreement with previous knowledge.

Table 6.1: **Set selection criteria and size of resulting gene sets**
Note: Gene set sizes in this table refer to each single dataset. In later overrepresentation analyses only those genes are included which were detected in both datasets.

criterium	**Macrophages** (median>=6.5)			**Liver** (median>=7.0)		
for set selection	SFG	LFG	BG	SFG	LFG	BG
p-value	<=0.02	<=0.05	>0.05	<=0.001	<=0.01	>0.01
PCCd1d2	>=0.81	>=0.75	<0.69	>=0.66	>=0.60	<0.54
signal strength	>=0.44	>=0.37	<0.33	>=0.81	>=0.75	<0.69
selected genes	337	622	6544	428	701	3595
proportion of expressed genes	3.1%	5.7%	60.2%	4.8%	7.9%	40.6%
permuted profiles	8	23	9391	0	0	8292
proportion of selected genes (FDR)	2.4%	3.7%		0%	0%	

6.8 How sizes of circadian gene subgroups differ from expectations

Promoter properties are variably distributed among gene categories How are the two different types of promoters (LCP, HCP) distributed among the four tissue-specific gene categories (C, I, N, O) of the two cell types (mouse macrophages and liver cells)? To consider the impact of promoter properties on tissue-specific gene expression, gene categories are additionally classified according to the promoters' GC and CpG content to LCP and HCP genes as described in section 5.5. In a comparison of the proportions of promoter property classes in gene sets of commonly expressed (EE), commonly non-expressed (OO) as well as tissue-specifically expressed genes of macrophages (MT) and liver cells (LT) the tissue-specificity ascribed to LCP genes is confirmed (figure 6.6). Furthermore, the proportion of promoter properties varies among gene categories. This suggests a regulatory impact of promoter properties on gene expression patterns. For that reason promoter properties and expression levels are controlled in the overrepresentation study as described in section 5.6.

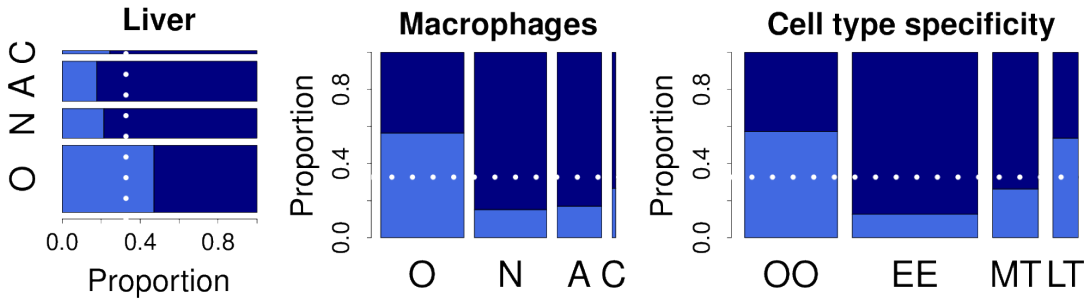


Figure 6.6: **CpG depleted genes are more abundant cell type specifically.** The barplots show relative sizes of gene categories as bar widths along with their proportions of LCP and HCP genes in *lightblue* and *darkblue*, respectively. Gene categories are circadian - C, ambiguous - A, non-circadian - N, non-expressed - O; Cell type specific groups are OO - non-expressed, EE - expressed in both cell types and expressed specifically in macrophages - MT or liver - LT. The *white dotted lines* mark the content of LCP genes in all detected genes.

Circadian genes of macrophages and liver overlap significantly Using the described time series analysis, 219 genes in macrophages (C^{Mac}) and 301 genes in liver cells (C^{Liv}) are classified as circadian out of 7152 genes, which were detected as expressed in both cell types (EE). How many genes do they have in common? Which size of overlap would be expected by chance if liver and macrophage cells did not share a basic core clock mechanism? A comparison between observed and expected overlap sizes shows the impact of the common clock mechanism on gene regulation. To calculate the expected overlap size, let us assume, that the proportion of macrophage circadian genes is the

same in all liver categories and vice versa. In this case, the expected overlap of circadian genes expressed in both cell types has the highest probability in the hypergeometric distribution. However, instead of the expected nine genes, we observe 44 common circadian genes. This is almost five times as much and according to the one-sided Fisher's exact test with a p-value of $7.2 * 10^{-19}$ highly significant (figure 6.7A).

With this, I bring in another view on the number of common circadian genes. Storch et al. [2002] perceived the overlap of genes, which were detected as circadian expressed in liver and heart, as "small" without giving any details on how large they expected the overlap to be. Their data show an overlap of circadian genes which is almost two times larger than randomly expected. These impressive overlaps of two datasets in each case support the idea that genes of different cell types share common regulatory features with respect to the circadian clock. Indeed, known core clock genes like *Per*s, *Cry*s, *Bmal1* and *RevErbα* are part of the common circadian gene set.

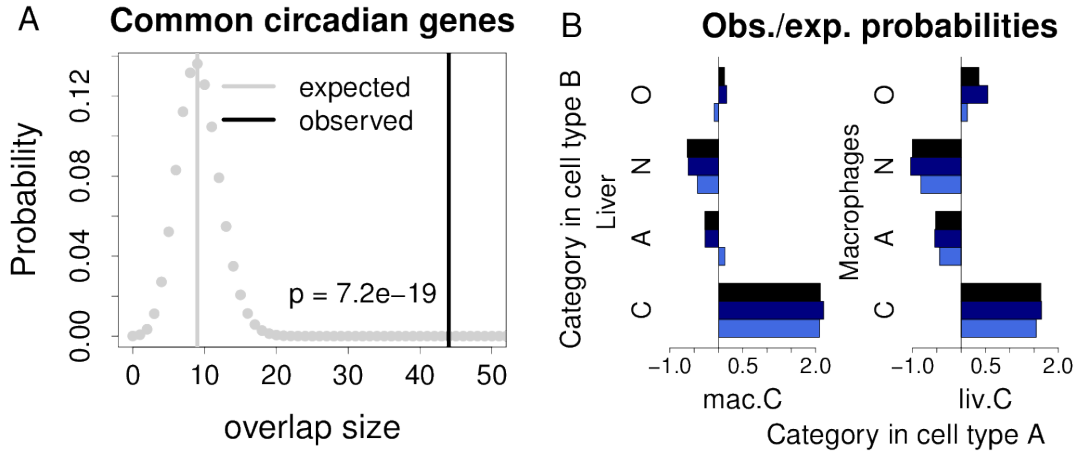


Figure 6.7: **Comparison of observed to expected sizes of category overlaps.** **A** The significant overlap of 44 macrophage and liver circadian genes underlines the presence of a common mechanism for circadian gene regulation. Random drawings of 219 balls (C^{Mac}) out of a pool of 7152 (EE) balls, of which 301 (C^{Liv}) are white, lead to the shown hypergeometric distribution of white balls among the drawn set. Most probably nine white balls are drawn (*grey line*); the observed CC set size of 44 is marked in *black*. **B** Ratios of observed to expected probabilities vary among categories especially if genes of interest are observed circadian in an other tissue. Ratios are shown on a logarithmic scale to base two, so that positve and negative bars indicate larger and smaller category comparison overlaps than expected. Promoter property classes are color-coded: LCP in *lightblue*, HCP in *darkblue* and ACP in *black*.

Differentially regulated genes are underrepresented If the core clock genes did not interact with tissue-specific transcription factors to regulate gene expression, the proportions of circadian genes of cell type A in all expression categories cat of the other cell type B were expected to equal the overall proportion of circadian genes $P(C_A)$. However, in reality these conditional probabilities $P(C_A|cat_B)$ vary. To calculate the probabilities $P(cat_B|C_A)$ of genes being classified as circadian (C), ambiguous (A), non-circadian (N) or non-expressed (O) in cell type B providing their transcripts are observed cycling in cell type A is calculated using Bayes' formula:

$$\begin{aligned} P(cat_B|C_A) = & \quad \frac{P(C_A|cat_B)P(cat_B)}{P(C_A|cat_B)P(cat_B)+P(C_A|cat_B^c)P(cat_B^c)}, \text{, where} \\ P(cat_B^c) = & \quad 1 - P(cat_B) \end{aligned} \tag{6.8}$$

Comparing observed to expected conditional probabilities in all classes of promoter properties reveals that common circadian gene expression is greatly enhanced by the clock mechanism while differentially regulated genes are underrepresented (figure 6.7B). In contrast to that, ambiguous genes of one cell type can be found in all categories of the other cell type with near expected probabilities. From this I conclude that circadian expression is favored over tissue-specific regulation. However, differentially regulated circadian genes may result from tissue-specific regulatory interactions of transcription factors that affect the oscillation range.

7 Overrepresentation analysis provides insights into gene regulation

Co-expressed genes are often targets of the same transcription factors and therefore share the presence of specific binding sites within their promoter regions. Based on this observation, regulators of co-expressed genes can be predicted using overrepresentation analysis, that compares the affinity distributions of transcription factors between these genes and a group of background genes. Thereby the transcription factors are represented by tabular descriptions of their binding motifs. Background genes are randomly chosen, more specifically, they are assumed not to share the similarity of expression patterns in the foreground group. Motifs with significantly higher affinities in foreground than background genes point to the importance of their binding factors for the foreground gene groups' expression regulation.

Circadian genes are characterized by a rhythm in expression level with a period of about 24 hours. Besides that they differ in phase timing and tissue-specificity of their rhythmicity. To be able to compare co-expressed circadian genes peaking within certain phase intervals with a proper background, all genes detectable by the two microarrays used in this work are grouped by their expression timing and the tissue-specificity of their expression and (non-)rhythmicity. Furthermore, foreground and background genes are classified by their promoter properties, because regulation dynamics differ between promoters with high and low nCpG content. To identify transcription factors associated with circadian regulation of expression or expression modulation, TF-promoter affinities are compared between subgroups of circadian and non-expressed or non-circadian genes chosen to match the other variables, according to the biological question. A hierarchical multiple testing method is applied to identify motifs with significantly higher affinities in foreground subgroups. Details for this method are given in section 7.1. Following that the importance of background choice is discussed. With regard to this issue, the relation of the chosen background to the biological question (section 7.3) and the impact of the matching procedure on the results (section 7.2) are considered. Finally, the overrepresentation results of the four comparison modes are put into context (section 7.4).

7.1 Hierarchical false discovery rate procedure reveals significant findings

Among thousands of tests comparing the affinity distributions for 611 motifs from TRANSFAC and Rey et al. [2011] between foreground and background gene sets, results for motifs involved in regulating circadian genes are expected to be enriched with signif-

icant p-values. However, each motif is tested under multiple conditions offering different chances for its overrepresentation: (1) the affinity distribution for a motif depends on the promoter class of the genes of interest; (2) if a motif is engaged for binding of transcription factors during a certain range of time only, chances are lower to find it in promoters of genes whose expression levels peak during the rest of the day; (3) a motif targeted primarily for tissue-specific gene regulation might be found more rarely in promoters of genes that are commonly expressed in the two cell types of interest. Thus, negative test results are expected for each motif, also for those important for circadian regulation. A method is needed that finds significant results like "needles in a haystack". This is possible with the "hierarchical false discovery rate-controlling methodology" described in Yekutieli et al. [2006], Yekutieli [2008] and Benjamini and Bogomolov [2011].

To find candidate motifs with a controlled FDR at significance level α=0.03, all hypotheses are grouped into disjoint subfamilies in a hierarchical tree as illustrated in figure 7.1. Hypotheses that belong to the same cell type, background choice, promoter property class, motif and phase group or category combination group are hierarchically tested by the Benjamini and Hochberg FDR-controlling (BH) procedure.

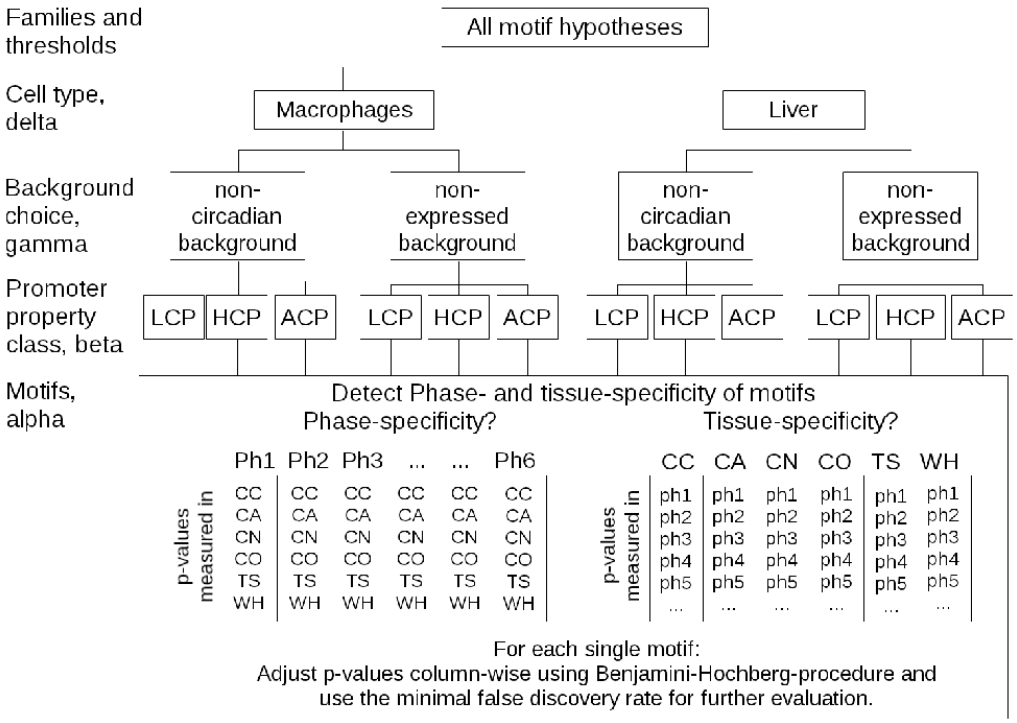


Figure 7.1: **Hierarchical tree of motif hypotheses.**
At each node the rate of falsely discovered motifs falls below the significance threshold α=0.03. FDRs for the phase- and tissue-specificity of each single motif are calulated in all promoter classes, background choices and cell types. Only motifs with FDRs below the thresholds indicated at each node are selected (for values see text 7.1).

Example Let's look at the double Ebox described by Rey et al. [2011]. Binding of the transcription factor BMAL1 has been shown to be enriched at promoters of circadian genes in liver, especially those circadianly expressed in several tissues. Can I support this statement at a significance level of α=0.03? Of all 432 p-values measured for the double Ebox in liver, ten fall below this threshold. All of the tested gene set pairs resulting in these p-values contained LCP genes. Their foregroud sets were chosen from circadian genes common to mouse liver and macrophages, their background genes from non-circadian liver genes which oscillate when expressed in macrophages. The 24 liver phase groups sourced from the foreground's category combination group (CC.LCP) form one of six category combination families. All 24 p-values within this one family are BH-adjusted. One of them falls below the threshold α=0.03, so that the FDR for the double Ebox to be enriched in common circadian genes can be estimated to $0.03 * (1 + 6)/(1 + 1)$=0.105 (figure 7.2A, see section 4.4.2). In contrast, as no family members of the other category combination groups contain significant BH-adjusted p-values, the FDR to find the double Ebox overrepresented in tissue-specificly regulated or expressed circadian genes is larger with $0.03 * (0 + 6)/(0 + 1)$=0.18. This is in line with the finding that genes bound at double Eboxes by BMAL1 are preferably common circadian genes.

To find out the timing of the double Ebox' overrepresentation, we look for the phase family with minimal FDR. Within one promoter class (in this case LCP) each of the 24 phase groups contains p-values of six category comparison groups, which are separately BH-adjusted. Now only at CT10 one BH-adjusted p-value falls below the treshold α=0.03. Thus this phase group's FDR is $0.03 * (1 + 24)/(1 + 1)$=0.375, while finding the double Ebox enriched in the other phase groups is marked with a larger FDR of 0.72 (figure 7.2B). This finding supports the report that BMAL1 target genes peak at CT10 (Rey et al. [2011]).

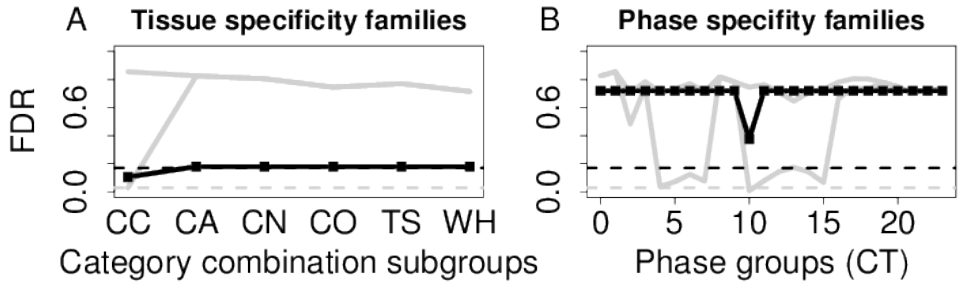


Figure 7.2: **The tandem Ebox is enriched in common circadian LCP genes peaking at CT10 in liver.**
The analysis described in the text is shown. For simplicity, only the range of BH-adjusted p-values within a family is marked with grey lines. The number of those values falling below the significance threshold $\alpha = 0.03$ (*dashed grey*) is used to estimate each family's FDR (*black squares*). A motif is overrepresented in circadian genes of a phase or category combination group, if this FDR is smaller than the cutoff β (*dashed black*).

Method details For each background choice (bg) in each cell type (tis) candidate motifs are selected according to their false discovery rate for being phase- and tissue-specific. Considering each promoter property class (ppc) separately, all p-values measured for one motif mtf form families according to the number f of phase (phs) or category comparison sets (ccs), within which the member p-values are BH-adjusted. To estimate each motif's FDR with respect to its phase- or tissue-specificity, the numbers d of the BH-adjusted p-values falling below the threshold α=0.03 in each family (fam) are used in the following formula:

$$FDR_{\alpha}^{mtf}(fam) \approx \left\{ \begin{array}{rl} \alpha * \frac{d+f}{d+1} & \text{if} \quad \alpha * \frac{d+f}{d+1} \leq \beta_{tis}^{fam} \\ 1 - \alpha * \frac{d+f}{d+1} & \text{if} \quad \alpha * \frac{d+f}{d+1} > \beta_{tis}^{fam} \end{array} \right\} \tag{7.1}$$

It is controlled at level $\beta_{tis}^{fam} = \alpha * (d_{min} + f)/(d_{min} + 1)$ with d_{min} being one sixth of the family members. Hence, values are $\beta_{mac}^{phs} = \beta_{mac}^{ccs}$=0.105 in macrophages and β_{liv}^{phs}=0.375, β_{liv}^{ccs}=0.105 in liver cells. A motif in the intersection of candidates for both criteria with respect to a particular promoter property class (ppc) has the false discovery rate:

$$FDR_{\alpha}^{mtf}(ppc) = FDR_{\alpha}^{mtf}(phs) + FDR_{\alpha}^{mtf}(ccs) - FDR_{\alpha}^{mtf}(phs) * FDR_{\alpha}^{mtf}(ccs) \tag{7.2}$$

The probability of a motif's false detection in any promoter class of one comparison mode, comprised of cell type (tis) and background choice (bg), is controlled at level $\gamma_{tis} = \beta_{tis}^{phs} + \beta_{tis}^{ccs} - \beta_{tis}^{phs} * \beta_{tis}^{ccs}$ (γ_{mac}=0.20, γ_{liv}=0.41):

$$FDR_{\alpha}^{mtf}(tis, bg) = \left\{ \begin{array}{r} FDR_{\alpha}^{mtf}(LCP) * FDR_{\alpha}^{mtf}(HCP) * FDR_{\alpha}^{mtf}(ACP) \\ \text{if} <= \gamma_{tis} \\ 1 - FDR_{\alpha}^{mtf}(LCP) * FDR_{\alpha}^{mtf}(HCP) * FDR_{\alpha}^{mtf}(ACP) \\ \text{in the other case} \end{array} \right\} \tag{7.3}$$

Finally, the FDR of a motif to be detected at least once in the whole promoter analysis by Wilcoxon test (W) is the product of the FDRs for the four comparison modes:

$$FDR_{\alpha,W}^{mtf} = FDR_{\alpha}^{mtf}(mac, nc) * FDR_{\alpha}^{mtf}(mac, ne) * FDR_{\alpha}^{mtf}(liv, nc) * FDR_{\alpha}^{mtf}(liv, ne). \tag{7.4}$$

If it falls below the threshold $\delta = (1 - \gamma_{mac})^2 * (1 - \gamma_{liv})^2$=0.22, the number of Wilcoxon p-values detected below the threshold α=0.03 for the corresponding motif is significantly high. However, the decision that a motif is found overrepresented in each single comparison mode may be false. The $FDR_{\alpha,S}^{mtf}$ for a motif to be truely detected in the observed

combination of comparison modes is:

$$FDR^{mtf}_{\alpha,S} = 1 - (p^{mtf}_{\alpha}(mac,nc)) * (p^{mtf}_{\alpha}(mac,ne)) * (p^{mtf}_{\alpha}(liv,nc)) * (p^{mtf}_{\alpha}(liv,ne))$$
$$\text{with } p^{mtf}_{\alpha}(tis,bg) = \left\{ \begin{array}{lll} 1 - FDR^{mtf}_{\alpha}(tis,bg) & \text{if} & FDR^{mtf}_{\alpha}(tis,bg) <= \gamma_{tis} \\ FDR^{mtf}_{\alpha}(tis,bg) & \text{if} & FDR^{mtf}_{\alpha}(tis,bg) > \gamma_{tis} \end{array} \right\} \quad (7.5)$$

The more generally a motif is involved in circadian gene regulation, the smaller gets this false discovery rate. For candidate motifs $FDR^{mtf}_{\alpha,S}$ falls below the threshold $\epsilon = 1 - (1 - \gamma_{mac}) * (1 - \gamma_{liv}) = 0.53$.

Two tables in the appendix list for all detected motifs ($FDR^{mtf}_{\alpha,W} \leq \delta$) the false discovery rates for each comparison mode ($FDR^{mtf}_{\alpha}(tis,bg)$) as well as the rates for overall ($FDR^{mtf}_{\alpha,W}$) and mode combination ($FDR^{mtf}_{\alpha,S}$) false discovery: Appendix A refers to motifs with $FDR^{mtf}_{\alpha,S} \leq \epsilon$ and appendix B to those with $FDR^{mtf}_{\alpha,S} > \epsilon$.

7.2 Impact of background matching on overrepresentation results

Overrepresentation analyses carried out in this study depend on many parameters. Some influence gene assignment to the circadian and non-circadian class and thereby influence the biological quality of the results (cutoffs for median, p-value for rhythmicity, day1-to-day2 correlation and signal strength, section 5.1). Likewise, the division of gene sets into LCP and HCP genes is rather a biological question on the influence of the general promoter sequence structure on gene regulation (section 5.5). Other statistical parameters determine the kind of background matching: the maximal distance of matching background genes that are allowed for each foreground gene determines the range of GC- and nCpG-content which is accepted as "similar". In case a smaller similarity range is preferred, there are less matching genes available to calculate a reliable statistic. Sampling more background sets leads to multiple use of several genes. Although the parameters are chosen carefully, such arbitrary selections affect the outcome of analyses and may introduce errors. In this section I evaluate the impact of the technical parameters on overrepresentation results.

7.2.1 Robustness of p-values

The variable technical parameters of background matching are the maximal distance of matching background genes allowed for each foreground gene and the total number of background sets to be produced. The latter one is a trade-off between calculation time and precision, while the first one might influence the accuracy.

To assess the effect of the matching distance on overrepresentation results, two foreground-background set pairs are selected for exemplary overrepresentation analysis using various sizes of matching distance. The first foreground set differs significantly from its background set with respect to GC and nCpG content as detected by Kolmogorov-Smirnov-test, so that matching would be desirable (common circadian and common

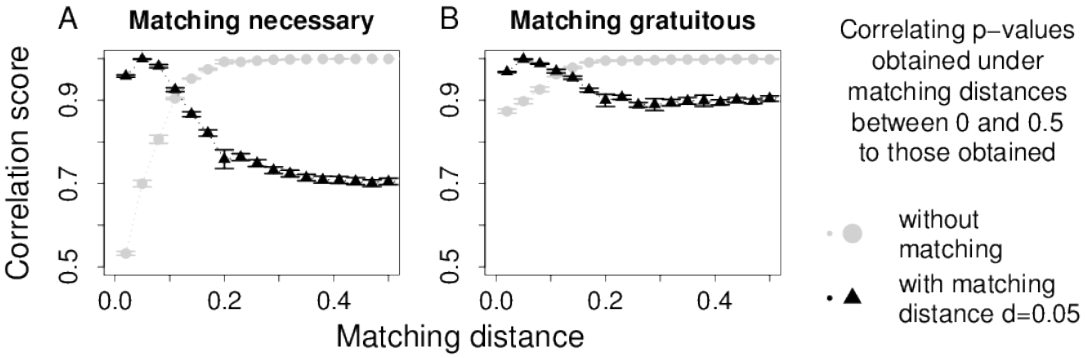


Figure 7.3: **Matching is needed and helps to exclude false discoveries.** Two set pairs are analyzed with varying parameters for the matching distance: for one matching is necessary (**B**), for the other one it is gratuitous (**C**). Overrepresentation results calculated without matching (*grey*) and with matching distance d=0.05 as used throughout this work (*black*) are compared by Pearson correlation to the outcomes of analyses using a range of matching distances between 0.02 and 0.5. By comparison, the robustness of overrepresentation results depends on the similarity of foreground and background sets with regard to promoter property distributions. Each overrepresentation analysis is carried out in triplets, so that error bars represent standard deviations of nine Pearson correlations.

non-circadian genes, $p_{KS}(GC) < 0.05$, $p_{KS}(nCpG) < 0.05$). For the second set pair the matching procedure would be gratuitous, because the comparison of the promoter property distributions does not justify the necessity of matching (liver-specifically expressed circadian and non-circadian LCP genes, $p_{KS}(GC) > 0.05$, $p_{KS}(nCpG) > 0.05$). Affinity distributions for 611 motifs are compared between foreground and background gene sets in three replicates of background sampling. To characterize the influence of matching on overrepresentation results, Pearson correlation is applied. Therefore, p-values calculated without matching or with matching distance d=0.05 are compared to sets of p-values calculated with matching distances between 0.02 and 0.5 (figure 7.3).

Apparently, results obtained with matching differ from results obtained without matching. However, the correlation is overall better with PCC≥0.9, if the matching was gratuitous. In the other case the correlation coefficient declines to 0.7 when comparing overrepresentation results obtained with extremely different matching distances. This indicates that differences in promoter property distributions between fore- and background impact promoter analysis in certain set pairs and applying a matching procedure during the background choice is a necessary control to decrease the number of false discoveries.

The disadvantage of very sharp matching (d=0.02) is the low number of available matching genes for some foreground genes. To be able to select background genes from a pool of at least 10 matching genes per foreground gene, I decided to use the matching distance of 0.05 in this study. This also correlates with the box width value of 0.1 chosen in Bozek et al. [2009].

7.2.2 Improved significance of motif prediction

In the last section we have seen that applying background gene matching affects motif enrichment results. How much does this procedure improve the prediction of motifs involved in circadian regulation? To answer this question, contingency tables as shown in figure 5.8 compiled from motif predictions based on overrepresentation analyses with and without background matching (otherwise with the same parameters) are interpreted.

The number D of motifs detected with significantly higher affinities in circadian genes is two to three times larger when no background matching is applied. This higher sensitivity (TP/m_1) of the analysis without matching is counterbalanced by an increased specificity when matching is applied: In this case the proportion of truely negative (TN) hypothesis tests among all null hypotheses (m_0) is larger (TN/m_0). This effect is more pronounced in macrophages, where the specificity grows from 55% without matching to 87% with matching distance d=0.05. In liver, specificity increases by 6% to 86%. Taken together, the significance - the proportion of falsely positive hypotheses (Wilcoxon p-values>α assigned to detected motifs) in all hypotheses with non-significant p-values (FP/m_0) - is decreased and therefore improved when matching is applied (figure 7.4).

Hence, despite a large false discovery rate for motifs involved in tissue- and phase-specific gene regulation, it seems that by using background matching motif prediction increased its ability to exclude motifs unrelated to circadian gene expression (the specificity) at the expense of its sensitivity.

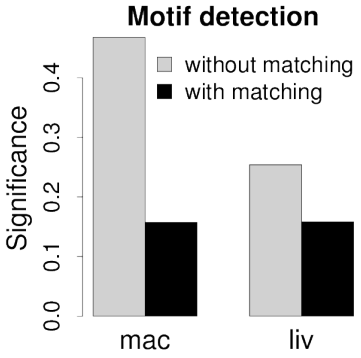


Figure 7.4:
The significance of motif prediction improves when applying a matching procedure for background choice. Wilcoxon test p-values below the significance level α=0.03 were used to predict motifs involved in the regulation of circadian genes in liver and macrophages controlling the FDR hierarchically. Significance levels in the final contingency table (see figure 5.8) are compared between analyses performed with or without background matching (d=0.05).

7.3 Background choice affects overrepresentation results

How does the more thorough background choice change previous experiences with overrepresentation analysis? To answer this question, the finding of certain Ebox motifs is tracked with different background choices. Separating genes by their promoter property classes reveals that the binding site for CLOCK:BMAL1 is overrepresented only in promoters with CpG islands (section 7.3.1). Moreover, in comparison to non-expressed genes the tandem Ebox identified for strong circadian binding of BMAL1 by Rey et al.

[2011] is enriched in all expressed genes, circadian and non-circadian ones (section 7.3.2). Furthermore, general Eboxes are more tissue-specifically found than canonical Eboxes (section 7.3.3). Finally, the kind of overrepresented motifs in circadian genes depends on the expression level of the chosen background genes (section 7.3.4).

7.3.1 Enrichment of canonical Ebox differs between promoter classes

Do common circadian genes differ in their binding affinities to the binding site of CLOCK:BMAL1 from genes, which do not show oscillations in both cell types? Following previous studies, all genes are defined as background genes that are detected by both microarray platforms and are not classified as common circadian. Affinity distributions to the motif CLOCKBMAL_Q6 with high information content in its "CACGTG" consensus sequence are compared using Wilcoxon test. In this setting, the chosen Ebox motif is clearly enriched with higher affinities ($p<0.05$, figure 7.5 ACP).

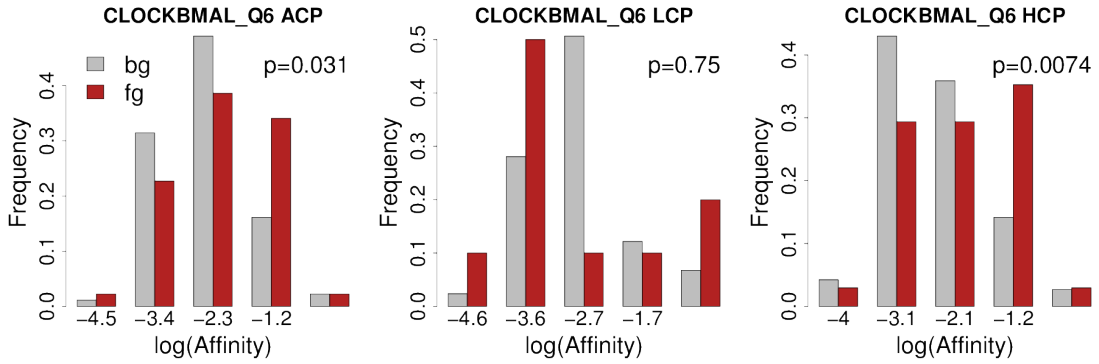


Figure 7.5: **Eboxes are enriched in circadian genes with high CpG content.** For the set of 44 common circadian genes (*fg*, 10 LCP and 34 HCP genes) a 500 times larger background set is chosen from the pool of all other 16426 genes detected by both microarray platforms analyzed (*bg*, 5372 LCP and 11054 HCP genes) to match the foreground's distributions of GC and nCpG content. Distributions of promoter affinities to the Ebox motif CLOCKBMAL_Q6 are compared by one-sided Wilcoxon rank sum test in three promoter property classes (ACP, LCP, HCP).

Knowing that the motifs' affinity distributions differ between genes with low and high CpG content (section 5.5), the question is, whether the significance is true in both subgroups. Repeating the test in LCP and HCP genes separately leads to the observation that only in the HCP gene class circadian genes differ significantly from this background choice ($p(HCP)<0.01$). As CLOCKBMAL_Q6 is a facilitator motif preferring binding to promoters with high GC and nCpG content, it is possible that cooperative binding of transcription factors preferring binding to specifier motifs is required in low CpG promoters to stabilize CLOCK:BMAL1 binding and elicit circadian expression.

7.3.2 Double Ebox is enriched in CpG-rich expressed genes

The paired binding of CLOCK:BMAL1 transcription factors to a double EBOX motif (EBOX_DBL) was shown to be highly predicitve for circadian oscillation of target genes (Rey et al. [2011]). Is this motif also overrepresented in promoters of circadian genes common to mouse liver and macrophages? Surprisingly, affinities of common circadian genes to this motif do not differ significantly in any promoter property class from those of background genes as defined before (p(ACP)=0.16, p(LCP)=0.36, p(HCP)=0.13).

However, when comparing the affinity distributions of common circadian genes to those of common non-circadianly expressed genes or genes non-expressed in both cell types, again the strongest differences are found in the group of genes with high CpG content. Promoters of circadian as well as non-circadian HCP genes have significantly higher affinities than non-expressed genes ($p<0.05$), while the contrast of affinity distributions is weaker when comparing the two subsets of expressed genes ($p>0.05$, figure 7.6). This may be explained by the importance of Eboxes for general transcription initiation as reported by Koike et al. [2012] and Martelot et al. [2012]. In this context fits that 83% of all BMAL1-bound sites fall near expressed genes (Rey et al. [2011]), whereas only 1 to 5% of unexpressed genes are similarly bound (Koike et al. [2012]).

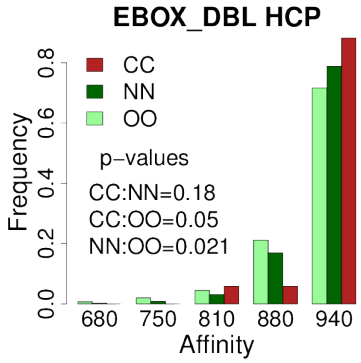


Figure 7.6:
Compared to non-expressed genes all expressed HCP genes are enriched with higher affinities to the double Ebox. The difference between circadian and non-circadian genes is not significant (Wilcoxon rank sum test). *CC*: 34 common circadian genes, *NN*: 1518 common non-circadian genes, *OO*: 2253 genes non-expressed in mouse liver and macrophages (all HCP), matching distance d=0.05.

7.3.3 Ebox motif variants may tissue-specifically influence rhythmicity

After focusing on genes with common pattern characteristics in mouse macrophages and liver cells, let's now compare the whole pattern categories in each cell type to learn about the tissue-specific differences in circadian gene regulation. Surprisingly, overrepresentation of the tandem Ebox is observed in the whole non-circadian categories of HCP genes in both cell types when comparing them to non-expressed genes (p(mac)=0.0022, p(liv)=0.042), while no significant differences between affinity distributions are detected between circadian and non-circadian categories in any promoter class. This may be due to the broad peak phase and tissue-specificity characteristics of these gene groups.

As peak phases of circadian genes are distributed over the whole day, it seems more appropriate to deem them co-expressed by phase group association (section 5.3). Plus,

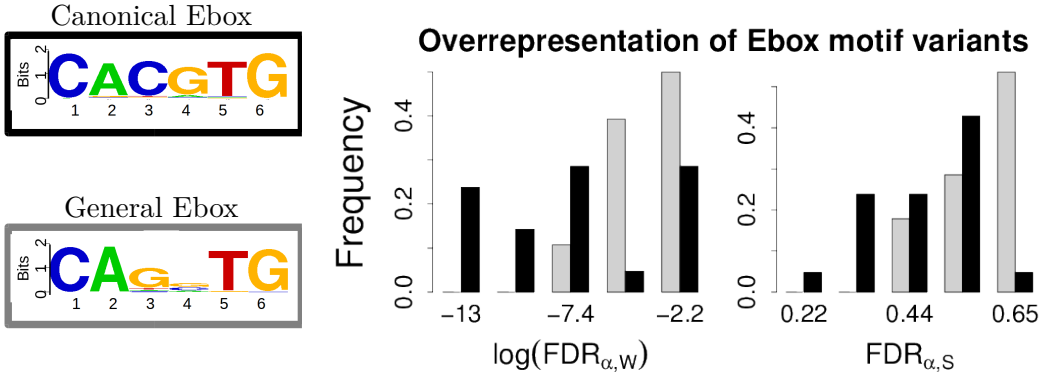


Figure 7.7: **Motif variants may determine cell type specificity of expression patterns.**
Basic helix-loop-helix transcription factors bind to two kinds of Eboxes, called canonical and general Eboxes (Consensus sequence logos with *black* and *grey* borders, respectively). Their overrepresentation is analyzed in various subgroup comparisons. Applying a hierarchical FDR control method to the p-values of Wilcoxon test yields an overall $FDR_{\alpha,W}$ for the significance of detection and another $FDR_{\alpha,S}$ for the number of detections in the four background comparison modes. The distributions of these false discovery rates for 20 canonical and 28 general Ebox motifs are shown (*black* and *grey* bars, respectively).

each non-circadian category contains genes circadianly expressed in the other cell type. To assemble co-expressed genes according to their tissue-specific rhythmicity, category comparison subgroups are formed (section 5.2). Overrepresentation analysis is now carried out comparing many subgroups of circadian and non-circadian or non-expressed genes in each promoter class. Unexpectedly, in this setting the tandem Ebox is not found significantly overrepresented any more.

However, the two Eboxes joined together by a linker of six to seven base pairs are bound by CLOCK:BMAL1 cooperatively. The information content differs among the two single motifs, the first one being more important for initial CLOCK:BMAL1 binding (Rey et al. [2011]). Based on this observation I ask, with which false discovery rates canonical ("CACGTG") and general ("CANNTG") Eboxes are found overrepresented in circadian genes using hierarchical FDR control of the many subset hypotheses tested. Clearly, in comparison to the general Eboxes with less information content canonical Eboxes are detected with smaller false discovery rates overall ($FDR_{\alpha,W}$) and among the four comparison modes using two background pools in each tissue ($FDR_{\alpha,S}$, figure 7.7). General Eboxes are found primarily in tissue-specific comparisons of circadian to non-expressed genes, while canonical Eboxes are detected in all kinds of comparison modes.

This suggests that binding site variants influence cell type specific regulation of circadian rhythmicity. This hypothesis is supported by findings of Badis et al. [2009], who showed that transcription factors with high affinities for the same octamer binding motif may be attracted with different strengths to lower-affinity binding sites. Furthermore, recent research in budding yeast has shown that motif variants "play an important role

in condition-specific gene regulation“ (Rest et al. [2012]).

7.3.4 The ratio of predicted motif types depends on background choice

How are motif variants distributed among gene groups to achieve tissue-specific circadian expression? In section 5.5 I showed already, that motif variants are sensitive for GC and nCpG content. These promoter properties are known to differ between tissue-specific and common expressed genes (section 6.8) as well as between non-expressed and expressed genes, while the two expressed categories of circadian and non-circadian genes are generally more similar (section 5.6.1). However, transcription factors choose their targets by scanning the accessible genome for their particular favourite binding site, thereby also crossing temporarily the sequences of many non-target genes (Hager et al. [2009]). Taken together, these observations suggest differences in the ratio of predicted motif types dependent on the kind of chosen background gene pool.

To go into the matter, all motif's affinity distributions are compared between circadian and non-circadian as well as circadian and non-expressed gene sets in mouse macrophages and liver cells, respectively (four comprison modes). The Wilcoxon p-values of all hypotheses tested in phase and category comparison subgroups of the three promoter classes are used in the hierarchical FDR control method to calculate false discovery rates for each motif. The proportions of specifier and facilitator motifs in the set of predicted motifs ($FDR_{\alpha}^{mtf}(tis, bg) <= \gamma_{tis}$) in each comparison mode are calculated. Interestingly, in both cell types the proportion of overrepresented specifier motifs is higher when comparing circadian to non-expressed genes. However, this change of ratio is only significant in liver (two-sided Fisher's exact test, p=0.0003) as shown in figure 7.8.

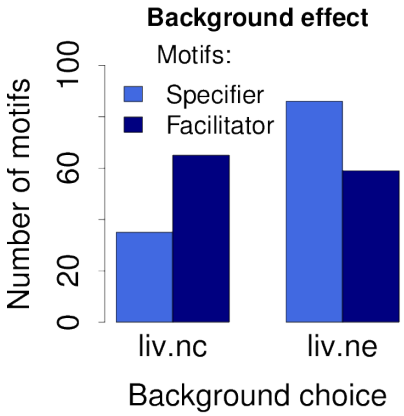


Figure 7.8:
Ratio of predicted motif types depends on background choice.
Overrepresentation analysis of subgroups is carried out for liver circadian genes using non-circadian (nc) and non-expressed (ne) genes as background pool separately. Motif prediction relies on the hierarchical FDR control method at significance threshold α=0.03 for Wilcoxon test p-values.

This result seems reasonable. We can assume that circadian and non-circadian genes expressed in the same cell type are both exposed to the common influence of tissue-specifically expressed transcription factors. Target binding sites of these "specifiers“ (Ravasi et al. [2010]) can be predicted from the comparison of circadian to non-expressed genes (see section 7.4). The preferential finding of specifier motifs by this comparison mode underlines the tissue-specific selectivity of transcription factors for genes with low

CpG content that need nucleosome removal for their accessibility. Once this is established by lineage-specific pioneering factors, binding of transcription factors preferring facilitator motifs is eased. As observed, circadian genes in liver offer significantly more high-affinity facilitator binding sites than non-circadian genes in their promoter regions (e.g. Eboxes). Tissue-specific expression timing may hence be determined by the timed interaction of transcription factors binding to specifier and facilitator motifs within each gene's promoter region. The more cell types a gene is expressed in, the less tissue-specific and the more ubiquitous transcription factors will bind its promoter, leading to better synchronization of the oscillating gene's expression profiles among several tissues. This explains the increased similarity between tissue-specific peak phases of individual genes cycling in a larger number of tissues as observed by Yan et al. [2008].

7.4 Enriched motifs provide a resource for prediction of TF interactions

Two biological questions are evaluated using time series data of mouse macrophages and liver: The first one asks very general for transcription factors that trigger circadian expression. They are predicted based on binding site affinity comparisons between promoters of expressed genes with circadian pattern and promoters of non-expressed genes while controlling for promoter properties and gene subgroupings (as described in chapter 5). However, since circadian genes differ in their promoter properties about as much from non-expressed genes as non-circadian genes do (section 5.6.1), we could assume that transcription factors found by this way are mainly responsible for tissue-specific expression and target circadian as well as non-circadian genes. To find regulation differences between the latter two groups, the second question focuses on the enrichment of binding sites for transcription factors, that modulate tissue-specifically the oscillation of expressed genes. Here, only promoters of expressed genes are used and those of circadian and non-circadian genes are compared while controlling for promoter properties and subgroupings. After answering these two questions in each cell type alone (four possible settings), a comparison of the results in a four-dimensional Venn diagram (figure 7.9) allows evaluating the impact of factors binding to the overrepresented motifs on tissue-specific circadian gene expression and expression modulation. Subsets of the candidate motifs represented in the Venn diagram are discussed in the following paragraphs. The biological involvement of regulators binding to these motifs in tissue-specific functions and their timing is exemplarily verified by literature.

7.4.1 Number of overrepresented binding sites

Based on the significance threshold α=0.03 almost half of all motifs used (302 out of 611) have been found in at least one of the four comparison modes. However, only 70 of them are significant when applying multiple testing control over all hypothesis tests in both cell types. Most tissue-specifically detected motifs have higher false discovery rates than commonly detected motifs (figure 7.9). The large number of motifs

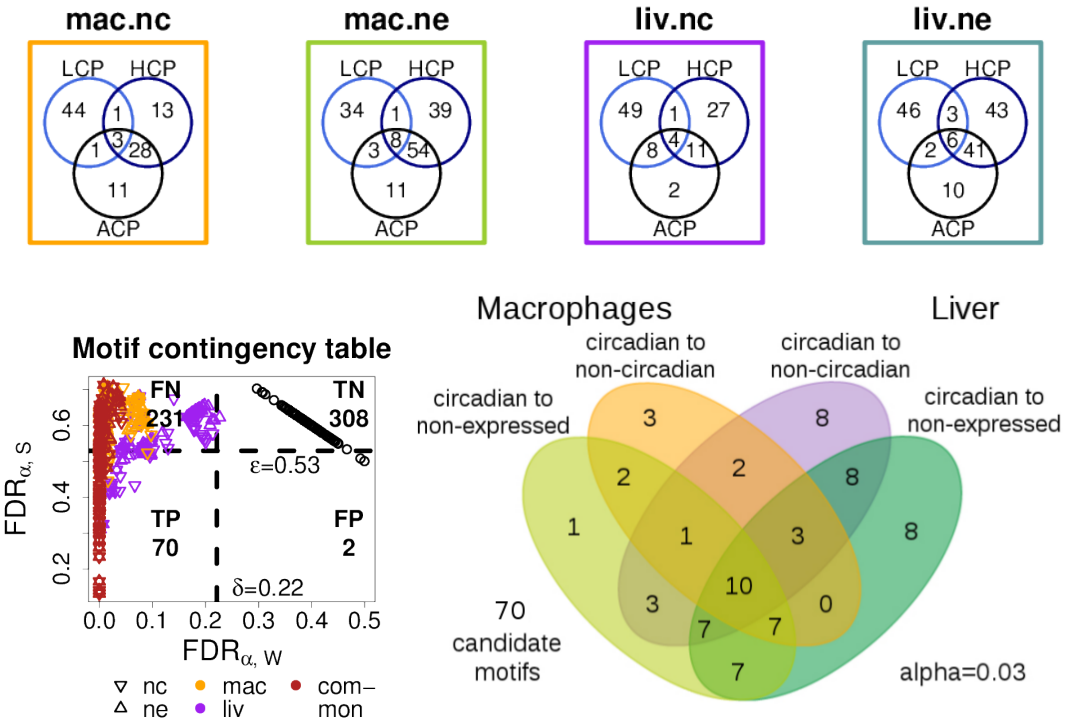


Figure 7.9: **Circadian genes share high binding affinities to common motifs. Upper row** Venn diagrams show the portions of 611 motifs detected as overrepresented in the indicated comparison mode (circadian versus non-circadian genes (nc) and circadian versus non-expressed genes (ne) in macrophages (mac) and liver cells (liv)) and all promoter classes (promoters with low (LCP) or high (HCP) GC and nCpG content and their union (ACP)). **Bottom left** The overall and mode combination false discovery rates ($FDR_{\alpha,W}$ and $FDR_{\alpha,S}$) serve to predict motifs involved in circadian gene regulation using the indicated thresholds. The tissue-specificity of the findings is highlighted graphically, showing that common circadian regulators have the lowest FDRs. **Bottom right** The four-dimensional Venn diagram shows the portions of 70 candidate motifs (TP) that are found significantly enriched in each mode of comparison.

involved indicates a more complex mechanism of tissue-specific circadian regulation than previously suggested. It is possible, that the timing of transcription factor interactions depends on the cell type and impacts effective transactivation. However, the Ebox of the form "CACGTG", Ybox ("CCAAT") and cAMP response element (CRE, "TGACG") are confirmed in their general importance for circadian genes, as they are found always overrepresented independent of the cell type or background mode chosen. These boxes are represented with ten motifs in the center of the four-dimensional Venn diagram (figure 7.9). Their binding factors are basic helix-loop-helix (BHLH), histone-like NF-Y and basic leucine zipper transcription (BZIP) factors, respectively.

7.4.2 Regulators for common expression and/or timing

Common circadian genes are important for many peripheral clocks: they generate rhythmicity, synchronize to circulating factors and control transcriptional outputs (Storch et al. [2002]). Despite the different nucleosome landscape and transcription factor sets present in each cell type expression of these genes reliably oscillates. Thus, motifs to which regulators for these genes bind are expected to be found overrepresented in both cell types targeted and by both comparison backgrounds. As already mentioned, this concept turned out to be true, as Eboxes ("CACGTG"), Yboxes and cAMP response elements are found in the intersection of the results in these four settings.

Motifs that distinguish circadian genes from non-circadian genes in both cell types are found in the upper subset on the central axis of the Venn diagram (two binding sites for MyoD and E2A). Regulators binding to these motifs would be expected to generate rhythmicity by modulating the timing of gene expression in both cell types. Consistent with this idea the motifs in this group are general Eboxes ("CANNTG"), potential low-affinity binding sites for CLOCK:BMAL1 and other more tissue-specific factors. E2A plays a major role in early B-cell differentiation and also binds within the insulin gene transcription control region (Crown Human Genome Center [2013]).

A list of seven motifs is found in both cell types enriched in circadian genes compared to non-expressed genes, but not in comparison of circadian to non-circadian genes. They are located in the lower subset on the central axis of the Venn diagram. Regulators binding to these sites are probably needed in both cell types for general gene expression and might be important for the timing of common functions of the two cell types. Indeed, according to the fact that macrophages and liver play important roles in the body's immunity, binding motifs for IRF are found in this subset besides three Eboxes for binding of MAX, USF and CMYC as well as th GC-rich motif SP1. These transcription factors are known to regulate genes involved in immune reactions.

Interestingly, also a GATA box was identified in all comparison modes, but its mode combination false discovery rate $FDR_{\alpha,S}$ is too large to let it emerge as candidate motif. This explains why its involvement in circadian regulation could not be experimentally verified (Schellenberg [2008]), after it was predicted before (Bozek et al. [2010]). GATA proteins may serve as transcriptional cofactors to select tissue-specific promoters, as GATA factors are important for cell differentiation of hematopoietic stem cells to macrophages (Ferreira et al. [2005]) and endo-/mesoderm to liver (Zheng et al. [2013]). The finding of a GATA motif as a result with low FDR suggests that cell differentiation and tissue-specific circadian gene selection are tightly interwoven.

7.4.3 Regulators providing tissue-specific information

A large proportion of circadian genes is tissue-specifically expressed or oscillates in only one of the cell types measured. Very few genes circadian in both cell types differ in their timing between macrophages and liver. This may be due to influences of additional transcription factors with either tissue-specific expression, activity or interaction patterns. Concerned genes might be regulated by transcription factor complexes with general and

tissue-specific cooperation partners according to the fuzzy puzzle model (section 2.2.2).

Binding sites for the common part of these transcription factor complexes should be found enriched in circadian genes of both cell types. In addition to the motifs for common regulators already mentioned, 21 motifs found in the six multi-overlaps in the central part of the Venn diagram are further putative candidates (four triple and two twofold overlaps including at least one comparison mode of each cell type). Besides further response elements for BHLH, BZIP and NF-Y transcription factors they contain motifs targeted by regulators with forkhead, zinc finger and helix-turn-helix domains. Their specific abilities to interact with cofactors, mediator or scaffold proteins enable them to recruit more tissue-specific transcription factors implicated in cell type differentiation. These interactions may be able to fine tune the peripheral clock's timing. Motifs overrepresented tissue-specifically are found in the six most peripheral subsets of the Venn diagram.

For macrophages, a binding site for MYB is found when comparing circadian to non-expressed genes. This transcription factor plays an essential role in the regulation of hematopoiesis (Crown Human Genome Center [2013]). Its name is derived from myeloblastosis, the differentiation pathway for macrophage progenitors. In comparison to non-circadian genes, binding sites for E2F, TATA-box binding protein and BHLH factors USF and DEC are found overrepresented.

In promoters of liver circadian genes motifs bound by hepatic nuclear factors like HNF1 (homeobox family, HOX) and HNF3 (forkhead family) are enriched compared to promoters of non-expressed genes. Other homeobox factors bind to motifs found overrepresented in contrast to non-circadian genes: PAX3, OCT4, ALX4, PITX2. Further enriched motifs suggest DBP, NFκB (p50), the basic helix-loop-helix proteins E12 and MYOGENIN as well as the zinc finger factors EGR4, HIC1, MAZ and WT1 as modulators of circadian expression pattern. The involvement of albumin gene D-site binding protein (DBP) in circadian regulation has long been documented along with its specificity for target promoters (Fonjallaz et al. [1996]). Recent literature highlights a role for DBP also in hematopoiesis (SethuNarayanan [2011]). NFκB was established two years ago as target of CLOCK regulation in the absence of BMAL1, which inhibits this interaction, establishing a molecular link between circadian clock and immune response mechanisms (Spengler et al. [2012]).

Table 7.1 lists binding factors to the 70 candidate motifs with low mode combination $FDR_{\alpha,S}$ discussed above. Many of the other detected binding sites with high $FDR_{\alpha,S}$ are overrepresented in only one or two comparison modes of one cell type. Regulators binding them are listed in table 7.2. A large portion of these binding sites are general Eboxes with the consensus ("CANNTG") bound by transcription factors with basic helix-loop-helix domains (BHLH) and CRE motifs bound by factors with basic leucine zipper domains (BZIP). This underlines the general importance of these motifs for circadian gene expression. However, small motif variations coupled with tissue-specific protein interactions may serve to derive tissue-specific timing.

Two ways of regulation interaction between general and tissue-specific regulators are conceivable: (1) The latter open chromatin loci as pioneers, making them accessible for more generally expressed regulators or (2) tissue-specific factors transactivate genes

only from cell type-specifically selected chromatin loci after many more were opened by pioneering general regulators. The finding of overrepresented double Eboxes in promoters of circadian as well as non-circadian genes when comparing them to promoters of non-expressed genes (section 7.3.2) and the fact that CLOCK:BMAL1 is expressed in many cell types suggest that the heterodimer might work as a pioneering factor. This hypothesis has recently been investigated and validated experimentally (Menet et al. [2014]).

Table 7.1: **Overview of regulators significantly involved in circadian gene expression in mouse macrophages and liver.**
Regulators binding to the 70 candidate motifs with $FDR_{\alpha,W} \leq \delta$ and $FDR_{\alpha,S} \leq \epsilon$ are sorted by their structure into families. For each family representatives are listed for which binding sites have been found overrepresented in macrophages (mac) or liver (liv) only or in both cell types (common). The font indicates whether they are involved in gene expression only (normal), in circadian modulation (*italic*) or in both (***bold and italic***).

TFclass	mac	common	liv
BHLH	*DEC, USF*	MAX, ***AHRHIF, ARNT, CLOCK:BMAL, MYC:MAX, MYC, SREBP1, STRA13, USF, USF2***, *E2A, MYOD*	***AHRARNT***, *E12, MYOGENIN*
BZIP		***CREB, TAX:CREB***	*DBP*
DM		DMRT1	
E2F	***E2F1DP2, E2F4DP2***	*E2F*	
FORKHEAD		*FOXO4, WHN*	HNF3
GRAINY			CP2
HISTONE		***ALPHACP1, CAAT, NF-Y***	***NF-Y***
HOX		***PAX3***	CHX10, HNF1, ***ALX4, OCT4, PAX3, PITX2***
HSF		***HSF1***	
IRF		IRF, IRF1	
MYB	MYB		
REL			EBF, *NFKAP-PAB50*
TBP	*MTATA*		
ZFC2H2		BLIMP1, CKROX, EGR2, GC, ROAZ, SP1, ZEC	***CACBINDING-PROTEIN***, *HIC1, MAZ, NGFIC, WT1*
ZFDOF			BARBIE

Table 7.2: **Overview of regulators with tissue-specific contributions to circadian gene expression in mouse macrophages or liver.**
Binding sites for these factors were found with $FDR_{\alpha,W} \leq \delta$ and $FDR_{\alpha,S} > \epsilon$. Sorting and font is configured as in table 7.1.

TFclass	mac	common	liv
BHLH	HEN1, TAL1BETA: E47, MYOD, TFE, ***E47***	***TAL1β: ITF2, SREBP1, TAL1***, *AP4*	ARNT, HAND1E:47, HIF1, ***TAL1α:E47***, *AP4, E2A*
BHSH	***AP2ALPHA***		*AP2*
BZIP	ATF1, ATF6, HLF, TAX:CREB, ***ATF4, CEBP***δ, *NFE2*	***AP1, ATF, ATF2, CREB, MAF***	ATF3, CEBP, E4BP4, VBP, *CHOP*
DM	DMRT7		DMRT4, DMRT5
ETS/ STAT	CETS168, ELK1, ETS, NRF2, PEA3, ***CETS1P54, GABP***	STAT3, ***PU1***	STAT1, ELF1
FORK-HEAD	***FOXM1***	FOXO1, FREAC4, HFH4, ***FREAC2***	FOXD3, FOXO4, FREAC3, HNF3
GRAINY	*LBP1*		*CP2*
HMG	***SOX***	SRY	***LEF1, SOX9***
HOX	PBX1, TGIF, *PAX8*	***CUT-like proteins, CRX, NKX25, XVENT1***, *HMX1*	HNF1, HOXA3, NANOG, OCT1, ***BRN2, HOXA7, MEF3, TTF1***, *IPF1, NKX25, PAX5*
IRF	*IRF7*	IRF2, ***ISRE***	ICSBP, IRF
MADS	SRF, *MEF2, RSRFC4*		*SRF*
REL	***CREL, NFKB p65, NFKB, RBPJK***	***NFKB***	NFAT
RUNT	AML1, PEBP2, ***OSF2***		
SMAD	***SMAD***	***NF1, SMAD3***	
ZFC2H2	GFI1, GZF1, NF-MUE1, SP3, SPZ1, ***IKAROS, MTF1, RP58, STAF, YY1, ZID***	***E4F1, SP1, ZBRK1***	IK3, YY1, *EGR3, EGR, MOVOB, SZF11*
ZFC4-NR (nuclear receptors)	DR3, LXR, PXR:RXR, ***DR4***, *ARP1, PPARG*	***androgen (AR) and glucocorticoid (GR) receptor***	COUP, FXR, HNF4ALPHA, HNF4, PR, RORA2, T3R, ***PPARA***, *SF1*
ZFGATA	*GATA3*	***GATA1***	*GATA2*

8 Discussion

This work set out to analyze promoters of circadian genes in mouse macrophages to predict transcription factors responsible for timing in their immune responses. Microarrays measuring gene expression at twelve timepoints during two consecutive days were the experimental base for this analysis (Keller et al. [2009]). In order to determine cell type specific circadian genes, a second microarray time series dataset of mouse liver with 48 timepoints was called in for comparison (Hughes et al. [2009]). The analysis of expression levels and patterns of genes that were detected by probesets on both microarray platforms confirms two previous observations: (1) promoters of tissue-specifically expressed genes tend to have lower GC- and nCpG-content than those of commonly expressed genes (6.8, Roider et al. [2009a]) and (2) circadian genes of a certain cell type contain besides a small part of genes found also circadian elsewhere a large part of genes tissue-specifically regulated or expressed in a circadian manner (6.8, Storch et al. [2002]). These observations are considered in more detail for predicting transcription factors involved in the regulation of tissue-specific circadian genes, thereby extending experiences with previous promoter analyses (Bozek et al. [2009], Yan et al. [2008]).

To characterize a gene with an expression profile, microrray probesets (the combination of several probes spotted on glass plates used to detect a certain DNA sequence) need to be assigned to genes. For the majority of genes this assignment is unique, only about 2-3% are detected by several probesets with partly conflicting characteristics. This observation may arise from post-transcriptional regulation of these gene products. As this is not the emphasis of my work, I decided to annotate only the probeset with the best circadian characteristics. This ensured not to miss out any circadian profile.

Based on the observation that transcription factors bind to short DNA sequences specifically, promoter analysis tries to predict transcription factors possibly involved in regulating the expression of a group of co-expressed genes by finding co-occurring binding sites within their promoter sequences. This concept was applied successfully to deduce factors whose activity had been modified in gene expression studies using microarrays. Thereby promoters of differentially regulated genes were compared to randomly chosen promoters of non-affected background genes to search for overrepresented binding sites (Meng et al. [2010]). Similarly, prediction of tissue-specific transcription factors is possible based on the enrichment of common genes at the top of two lists, one ranking genes by tissue-specificity and the other by their predicted affinity to each transcription factor (Roider et al. [2009b]). However, circadian genes are up- and downregulated at different times during the day and this pattern may even be tissue- specifically influenced. The choice of proper fore- and background gene sets is therefore an important step in promoter analysis of circadian genes. The following sections examine a list of questions linked to this decision and thereby highlight the results of this work.

8.1 What qualifies a gene as expressed versus non-expressed?

Gene expression measured by microarrays relies on DNA hybridization specificity. The small glass chip is spotted with synthesized DNA oligomers designed as reverse complements of as many as possible genes' coding DNAs. After hybridizing to fluorescence-marked cDNA amplified from isolated cellular mRNA the intensity of the color at each probe's spot is used to infer the associated gene's expression level. However, mismatched hybridizations and the sensitivity of the microscope influence the measurement results. These problems are considered when normalizing data with the GeneChip Robust Multiarray Averaging (GC-RMA) procedure, which results in a bimodal distribution of all genes' median expression levels. Control probesets targeting intronic and exonic regions of genes reveal the two underlying modes of noisy and biologically regulated expression, called here "non-expressed" and "expressed" (section 6.2). An expression cutoff was chosen manually to distinguish between these two classes. Notably, this cutoff cannot correctly reflect a biological discrimination between expressed and non-expressed modes.

From the biological view transcription factors are able to assemble and form a pre-initiation complex in the promoter region of an expressed gene that successfully recruits RNA polymerase. Thus, despite the possible presence and time-limited accessibility of random binding sites throughout the genome, expressed genes are able to attract more activating transcription factors than non-expressed genes over time.

8.2 What qualifies a gene as circadian versus non-circadian?

A circadian gene is transcriptionally activated by transcription factors interacting within a particular time frame while at other times its transcription is either repressed or attenuated. The combination of such a transcription profile with a short mRNA half life (<6 h) yields an observable circadian expression pattern (Jacobshagen et al. [2008]). It is expected to show regular fluctuations with a period of about 24 hours. Several methods have been used concurrently to determine such genes as accurately as possible (chapter 6): First, fitting a sine wave and comparing its residues to the ones of a constant fit by Ftest statistics resulted in a p-value (section 6.3). However, manual inspection of timeseries selected at the chosen significance level revealed many unconvincing examples. Often the wave fit was preferred because of a masurement outlier or was characterized by a low amplitude. To exclude such erroneous selections, two more criteria were applied: (1) Comparing the measurements of the two days with each other resulted in a Pearson correlation coefficient (section 6.4) and (2) the difference between peak and trough in relation to the base expression level resulted in a measure for the signal strength (section 6.5). In comparison to permuted timeseries it could be shown, that this triple approach reduced the false discovery rate (section 6.6). All three measures together yielded a circadian gene set that passed manual inspection of the measured time series. In support of the selection criteria many known circadian genes were detected as such (section 6.7).

However, this method may overlook circadian regulated genes for several reasons. The

circadian transcription of some genes seemingly non-circadian may be occluded by a very long mRNA lifetime and may only be revealed under changed degradation kinetics in an other cell type. Another possibility to overlook circadian regulation is when a gene seems non-circadian, but is regulated rotationally by two antiphase transcription factors. This has been shown for genes with 12 hour periods (Westermark and Herzel [2013]).

The detection of non-circadian expression profiles is less straightforward. They are filtered in this work by excluding all expressed genes that fullfill any very broad criterium for circadian expression. Based on this method it cannot be concluded with certainty, that there is no circadian transcriptional regulation for a gene whose expression pattern does not show a 24-hour period. However, in contrast to the previous use of the whole database, selection of background genes from microarray data is a progress. It could be improved by relying on nascent RNA instead of mRNA profiling to observe transcription products before posttranscriptional modifications take place (Martelot et al. [2012]).

8.3 How many common circadian genes are expected?

The proportion of circadian genes in a tissue, whose expression levels are also observed oscillating in other tissues, is perceived as "small" (Storch et al. [2002]). Based on the common core clock mechanism active in almost all cells displaying circadian rhythmicity expectations seemed to be higher. I calculated the expected size of the overlap between macrophage and liver circadian genes, if their rhythm was regulated only tissue-specifically (section 6.8). It is about five times smaller than observed supporting the strong influence of common mechanisms of circadian gene regulation among different cell types. These are rather based on promoter properties than on binding sites alone which will be discussed in the following sections.

8.4 What is a proper background gene set?

To answer a biological question on gene-regulation of co-expressed genes an appropriate background choice is necessary for overrepresentation analysis. The aspect of interest needs to contrast between fore- and background sets, while other aspects must be controlled for similarity. This thought led to the subgrouping of all genes detected by the two microarray platforms that were used for time series measurement in mouse macrophages and liver cells with respect to the aspects discussed hereafter (chapter 5).

(1) Circadian genes are expressed. In contrast to other expressed genes they show an oscillating pattern with a 24-hour period. These two characteristics call for two opposites: non-expressed genes, whose patterns resemble noise, and expressed non-circadian genes, whose patterns lack the rhythm, but are expressed as well. These two background sets differ with regard to their GC and nCpG content distributions (section 5.6.1). Comparing circadian genes to both of them separately helps to disentangle the related but distinct questions for transcription factors that regulate circadian expression or modulate patterns of expressed genes (section 7.3.4). To compensate for distribution differences between compared gene sets, background genes are chosen randomly from an "area"

around each foreground gene to create a gene set matching the GC and nCpG distributions of the foreground set (section 5.6.2). The application of this procedure leads to an increased specificity (true negative rate) of overrepresentation analysis (section 7.2.2).

(2) Many circadian genes oscillate tissue-specifically. Due to its known influence on tissue-specific gene expression, promoters with different properties regarding GC and nCpG content are analyzed separately (section 5.5). This yields more specific information than their mixed analysis (section 7.4.1). To increase the contrast with circadian and background genes in one cell type, both groups are chosen from genes categorized the same in the other cell type (sections 5.2 and 5.6). Interestingly, the binding profile of promoters for transcription factors targeting Ebox motifs seems to code for tissue-specific expression: The double Ebox is found overrepresented in common circadian and non-circadian compared to non-expressed genes (section 7.3.2), underlining the recent finding of pioneering activities of CLOCK:BMAL1 (Menet et al. [2014]). However, the false discovery rates of single Eboxes depend on their consensus sequences: While canonical Eboxes ("CACGTG") are found in both cell types, Eboxes of the type "CANNTG" are found mainly tissue-specifically (section 7.3.3). In similarity to the nucleotide change in the NFκB-motif, which regulates the binding transcription factor's dependence on a specific coregulator (Natoli [2004]), I hypothesize, that the Ebox binding profile offered by a promoter region affects both, the temporal and tissue-specific composition of the interacting transcription factor complex. Regulator binding and interaction does not depend solely on the presence of obvious binding sites, but also on the general sequence composition. The less tissue-specific transcription factors are attracted to a particular promoter, the more exclusively it is regulated by ubiquitous transcription factors such as CLOCK:BMAL1 leading to better synchronization of the gene's expression profile among several tissues (section 7.3.4).

(3) Circadian genes peak around the clock. Genes that are upregulated above their mean expression level during the same time interval can be viewed as co-regulated. A transcription factor active during a certain time interval searches for binding sites in the accessible sites of the whole genome, passing by circadian as well as all other genes. However, if it succeeds in overcoming repressing modes of promoter accessibility/occupation of a gene or in increasing transiently the kinetics of a gene's transactivation, the gene's expression might oscillate. To be able to compare in each phase interval promoters of upregulated circadian and other genes, all genes were sorted into phase groups (section 5.3). To control for expression timing, circadian and their background genes were compared within these phase groups. For the double Ebox the FDR for overrepresentation calculated at phase CT10 in liver cells was half the value of the FDR at all other phases (section 7.1). This timing coincides with the peak phase of cytosolic mRNA expression of BMAL1 targets in liver (Rey et al. [2011]).

8.5 What causes tissue-specific gene expression?

Mammalian organisms, starting from one single cell, develop many differentiated cell types throughout their body. All carry the same genomic information, but develop dif-

ferent characteristics with respect to gene expression and function. The tissue-specificity of a gene product increases, the less cell types it is expressed in at the same level or higher. Cell type-specific genes are regulated by transcription factor interactions between widespread expressed facilitators and tissue-restricted specifiers (Ravasi et al. [2010]). As promoters of widespread expressed genes are associated with high GC and nCpG content (HCP) and promoters of more specifically expressed genes with low GC and nCpG content (LCP, Valen and Sandelin [2011]), it is conceivable that facilitators are often but not exclusively expressed from HCP genes and specifiers from LCP genes. In the context of enriched LCP genes in tissue-specifically expressed genes (Roider et al. [2009b]) it is a fair question, whether transcription factors are able to distinguish between these two promoter types so that a regulator can have more tissue-specifically than generally expressed targets or the other way around. Comparing transcription factor affinities to each motif in promoters of the two classes revealed indeed two motif groups: "facilitator motifs", which attract TFs generally with higher affinities preferentially to HCPs and "specifier motifs" with a broad affinity spectrum including more LCP genes among the top ranks (section 5.5). Thus, besides the existence of appropriate binding sites general promoter properties influence the binding of transcription factors by providing a high or low affinity environment. Gordân et al. [2013] come to the same conclusion studying the in vivo and in vitro DNA binding specificities of two paralogous basic helix-loop-helix transcription factors targeting similar Eboxes using genomic-context protein binding microarrays.

In their search for binding sites transcription factors diffuse in three dimensions through the cell, but are also able to transiently bind to DNA (Hager et al. [2009]). In a high affinity environment they will spend more time than in a low affinity environment, increasing their chances to meet suitable interaction partners to regulate the targeted gene. If the interaction partner prefers a similar environment, they meet more often and wherever they are co-expressed. In the case of different preferences the transcription factor interaction depends on the ratio of their concentrations in the cell. This is variable between tissues, enabling different transcription factor interactions among tissues as observed by Ravasi et al. [2010]. To capture binding sites that are bound by transcription factors activating genes specifically in one cell type, the promoter comparisons between expressed and non-expressed, common and tissue-specifically expressed as well as tissue-specifically with nowhere expressed genes are useful (section 5.2). This concept is confirmed by the finding of tissue-specific transcription factors like MYB in macrophages and several HNF-factors in liver cells as well as common factors like CLOCK:BMAL1, MYC, CREB, NF-Y and IRF (table 7.1).

8.6 What causes tissue-specific oscillation of gene expression?

Expression oscillation is caused by feedback loops of transcriptional activation and repression with a delay. A feedback loop often implemented in natural oscillating systems is composed of activators, that may positively feed back on their own transcription, and repressors, whose transcription is initiated by the activators and who finally repress this

action after a time delay. The dynamics in such a system can be tuned by the availability of extra DNA target sites for the activator and/or repressor (Jayanthi and Vecchio [2012]). This impedance-like effect is called retroactivity and reveals "that a new, more subtle, role for the large number of inactive promoter sites on the chromosome capable of binding proteins (Burger et al. [2010]) is that of tuning the temporal dynamics of gene transcription" (Jayanthi et al. [2013]).

This is especially interesting with regard to the circadian clock and its tissue-specific gene regulation. In the datasets at hand 3-8% of genes showed cycling mRNA abundance levels (section 6.7), giving reason to the assumption that the concentration or histone acetylation activity of a certain percentage of transcription factors changed with time. However, with the changing concentration of a transcription factor or its histone acetylation activity the availability of binding sites alters with time. This feeds back on regulator binding kinetics at their target sites. As a consequence, the probability of a particular transcription factor interaction at a certain promoter depends on time and cell type.

Overrepresentation analysis and hierarchical testing on multiple tissue- and phase-specific circadian gene subsets detected binding sites for transcription factors that may play a role in such a mechanism (tables 7.1 and 7.2). Follow-up research could predict gene expression levels and patterns from calculated time-dependent probabilities for transcription factor interactions at circadian genes estimated based on promoter sequences, microarray time series and data of chromatin immunoprecipitation experiments. Transcription factor interactions predicted in this study could serve as a starting point. These ideas have already inspired the study of Korenčič et al. [2012], who found "that the intrinsic gene regulatory network primarily determines the circadian clock in liver, whereas systemic cues such as light-dark cycles serve to fine-tune the rhythms".

Examples for how transcription factors and their complexes with cofactors may influence gene expression timing differentially in different cell types are discussed in the following subsections.

8.6.1 Basic helix-loop-helix factors: binding canonical and general Eboxes

The basic helix-loop-helix (BHLH) superfamily constitutes a large class of transcription factors, which are partly ubiquitously expressed, while others are tissue-specifically expressed. They are involved in diverse functions: circadian clock, cell cycle, cell-lineage development and tumorigenesis (Jones [2004]). Proteins containing helix-loop-helix structures (HLH) are able to homo- or heterodimerize. If both components contain a basic region adjacent to the amphiphatic HLH domain, they are able to bind DNA (Guasconi et al. [2003]). Preferred binding motifs share the consensus "CANNTG", called the general Ebox (Kadesch [1993]). Additional domains in their protein structure (PAS: Per-ARNT-Sim domain, ZIP: leucine zipper) regulate the choice of binding partners, which impacts their final DNA binding specificity (Partch and Gardner [2010]). This was exemplary shown for the transcription factor AP4 (Hu et al. [1990]).

Many general Ebox motifs are found overrepresented in macrophage and liver circadian genes in both types of comparisons: against non-circadian and non-expressed genes.

However, in the overlap of these four result sets only motifs containing the "CACGTG" consensus are found, including the binding site for CLOCK:BMAL1 (section 7.4). This suggests that binding proteins to the canonical Ebox are general regulators of the circadian clock gene network and fullfill functions in many cell types, whereas binders to the general Ebox have rather tissue-specific functions. However, inside the nucleus of a certain cell general timing regulators may regulate the expression of or form transient complexes with more tissue-specific BHLH transcription factors forwarding the timing message to tissue-specifically regulated genes.

The MYC protein is an example for an ubiquitously active transcription factor. Its binding site is found overrepresented in all four comparison modes. The *Myc* gene is regulated by CLOCK:BMAL1 (Taniguchi et al. [2009]) and is involved in hematopoietic stem cell function (Laurenti et al. [2008]) as well as hepatic glycolysis and lipogenesis (Valera et al. [1995]). However, its expression profile is classified as circadian in the large foreground set of macrophages, but non-circadian in liver leading to the forwarding of different timing informations in the two cell types. As a secondary transcription factor MYC amplifies non-specifically a myriad of targets involved in many biological functions that are previously targeted by primary regulators (Nie et al. [2012]). It is also capable of repressing transcription of genes by blocking the activity of bound transactivators (Herkert and Eilers [2010]). Due to the global consequences of small changes in MYC levels its expression level must be precisely controlled (Levens [2010]): This includes dynamic binding and unbinding of multiple transcription factors. Their timing induces ratcheting the polymerase complex from initiation to promoter escape. This mechanism reduces intrinsic noise coming from cell-to-cell variation in MYC levels. Because multiple proteins bind the *Myc* promoter (Wierstra and Alves [2008]), it is conjecturable that direct or indirect target gene products feed back on its transcription. Due to its short half-life of about twenty minutes a "sequential change in coregulator composition at gene targets" has been proposed (Eilers and Eisenman [2008]). Moreover, other transcription factors also bind canonical Eboxes (Jones [2004]), thus competing with MYC for binding to these DNA sites. Post-translational modifications like phosphorylation, acetylation and ubiquitinylation affect MYC's promoter and binding partner selection, transactivation potential as well as its stability (Vervoorts et al. [2006]). Taken together, transcription factor expression, binding and interaction is dynamic at the input and output side of the *Myc* gene leading to different expression profiles in macrophages and liver.

TAL1 is an example for a tissue-specific basic helix-loop-helix (BHLH) transcription factor. It is involved in hematopoiesis leading to erythroid maturation (Kassouf et al. [2010]). Moreover, it enhances GATA1 binding to DNA, especially at combined GATA-Ebox sites including nine linker base pairs (Kassouf et al. [2010]). Based on the analyzed time series here, its gene is classified as expressed in macrophages and non-expressed in liver. Nevertheless its binding site, a kind of general Ebox with consensus "CAGATG", is found overrepresented in the circadian gene sets of both cell types when comparing them to non-circadian genes. While the finding in macrophages may have a functional background, finding it in liver may be due to the binding site similarity with other Eboxes found enriched in circadian genes or may hint on an interaction of Tal1 with different interaction partners.

8.6.2 Histone like factors: pioneering transcriptional activation

Nuclear factor Y (NF-Y) is a heterotrimeric transcription factor binding in proximal promoters as well as in distal enhancer regions. While its subunits B and C contact the DNA through a histone fold domain, subunit A confers sequence specificity for the consensus "CCAAT" (Ybox). Moreover, its knockdown seems to shift the ratio between proximal and distal promoter binding towards the latter one, which serves more tissue-specific regulation (Fleming et al. [2013]). The histone-like structure of NF-Y equips it with the ability to bind compacted chromatin lacking activating histone marks as a pioneering transcription factor (Fleming et al. [2013]). Its many targets are implicated in cell signalling, DNA repair, cell cycle, metabolism and gene expression, rendering NF-Y as a ubiquitous transcription factor.

Interestingly, in liver the gene *NfyA* coding for the A subunit is classified as non-expressed, although the binding site recognized by this subunit is overrepresented in all comparisons of circadian to other genes. As shown by Xiao et al. [2013], NF-YA knockdown leads to up-regulation of *Bmal1* promoter activity, and indeed, the BMAL1 expression level is higher in liver than in macrophages. Therefore, it may be able to activate more tissue-specific genes with rather low GC and nCpG content in a circadian manner. I suggest that the ratio between NF-Y and CLOCK:BMAL1 expressed in a cell influences the properties of gene expression patterns. The reported statistical enrichment of Eboxes near Yboxes argues for "a pervasive partnership" of these binding sites (Fleming et al. [2013]). Interactors of NF-Y include E2F, FOS, MYC, USF and other Ebox-binders, whose binding sites are also found enriched in circadian genes of mouse macrophages and liver cells.

Notably, several proteins in other transcription factor families are able to bind to the same sequence motif (Raymondjean et al. [1988]). One such potential binding competitor is the CCAAT/enhancer binding protein (C/EBP) containing a basic leucine zipper domain (BZIP), that is able to synergistically cooperate with NF-Y to activate genes (Xu et al. [2006], Shi et al. [2012]). An other protein is the CCAAT displacement protein (CDP/cut), a highly conserved homeodomain protein that acts as a transcriptional repressor (Li et al. [2000]). Binding sites for these proteins are found in both, macrophages and liver circadian genes, with rather tissue-specific character and high mode combination false discovery rate $FDR_{\alpha,S}$.

8.6.3 Basic leucine zipper factors: factors relaying extracellular stimulation

The most prominent transcription factor in the basic leucine zipper (bZIP) family is CREB, cAMP responsive element binding protein. Several binding sites for CREB and and its relatives CREM (cAMP response element modulator) and ATF-1 (activating transcription factor 1) with consensus "TGACGTAA" are found overrepresented in circadian genes in all kinds of comparison settings. CREB regulates the expression of genes involved in hematopoiesis (Sandoval et al. [2009]) and gluconeogenesis in the liver (Oh et al. [2013]).

Upon activation, CREB binds as second messenger together with its coregulators

CREB binding protein (CBP) or p300 to cAMP responsive elements (CRE) in the promoters of thousands of target genes to initiate transcription. Several protein kinases are able to activate CREB by phosphorylating its serine-133 residue. This happens upon extracellular binding of signalling molecules to receptors located at the cell surface which leads to elevated cyclic adenosine monophosphate (cAMP) or calcium-ion concentrations (primary messengers) in the cytoplasm (Nichols et al. [1992]). Such signalling ligands may be cytokines like granulocyte-macrophage-colony stimulating factor (GM-CSF) or glucagon in the cases of macrophage and liver cells, respectively (Sandoval et al. [2009], Oh et al. [2013]), which show circadian rhythms in their abundance levels (Akbulut et al. [1999], Gagliardino et al. [1978]). However, for proper gene activation CREB needs several regulatory partners; these coregulators are also able to achieve tissue-specificity in the cAMP-mediated cellular response to extracellular stimulation (Zhang et al. [2005]). Dependent its interaction partners' cofactors it is also able to modulate CLOCK:BMAL1-mediated transcription: it was shown that p300/CBP associate factor (pCAF) serves as coactivator and histone deacetylase 3 (HDAC3) as corepressor for CBP/p300-mediated modulation of CLOCK:BMAL1 transactivity (Hosoda et al. [2009]). Thus, via its dependence on several coregulators and its ability to modulate CLOCK:BMAL1 transactivation, CREB might serve as a mediator for the assembly of tissue-specific transcription factors at circadian genes. In support for this hypothesis, transcription factor binding motifs that have been found to co-segregate with cAMP responsive elements (Zhang et al. [2005]) were found overrepresented in circadian genes as well. These include Ebox and Ybox motifs found in both cell types as well as tissue-specifically enriched motifs for factors of REL, GATA and ETS families in macrophages and for AP2, E2F, NFY and USF factors in liver cells. This indicates indeed tissue-specific recruitment of cofactors.

8.6.4 GATA factors: Links between cell differentiation and the clock?

GATA transcription factors contain zinc fingers in their DNA-binding domain that bind specifically to the consensus site "WGATAR". These proteins are no core clock regulators, but binding sites for GATA-1 are found overrepresented in both, macrophage and liver circadian genes, compared to both, expressed non-circadian and non-expressed genes. Furthermore, as tissue-specific binding sites for liver and macrophages GATA-2 and GATA-3 are found, respectively, with high mode combination FDR when comparing circadian to non-circadian genes. How are GATA factors connected to the macrophage and liver gene expression as well as to the circadian clock?

Conserved throughout eucaryotes, GATA factors are involved in several developmental processes, especially in hematopoiesis, the maturation of blood cells of distinct lineages including macrophages (factors GATA1-3, Ferreira et al. [2005]) as well as endodermal and mesodermal tissues including liver (factors GATA4-6, Zheng et al. [2013]). Thus, they are important regulators in macrophages and liver besides other cell types and tissues. Their binding motifs were also found overrepresented in previous analyses of circadian genes (Bozek et al. [2010]). However, in subsequent experiments oscillation of luminescence could not be observed in U2-OS cells transfected with a luciferase reporter gene driven by a promoter containing repeated GATA binding sites (Schellenberg

[2008]). Is there really a connection between GATA factors and expression timing as again suggested by the actual overrepresentation results?

Knockout and rescue studies in erythroid cells showed that the dynamic spatiotemporal expression patterns of GATA(1-3) proteins are more important than their identities for developmental control (Ferreira et al. [2007]). These are encoded in their *cis*-regulatory elements and influenced by post-translational regulation. It is suggested that unique interactions of GATA factors with other semi-restricted transcription factors influence tissue-specific gene regulation (Molkentin [2000]). For instance, in Drosophila modules of the mediator protein complex CDK8 were shown to regulate the tissue-specific activity of the GATA/RUNX complex including the Gata factor SERPENT and the Runt box (RUNX) factor Lozenge (Gobert et al. [2010]). The Drosophila protein Lozenge plays a crucial role in cell fate decisions during the Drosophila eye development. Based on its homology to the mammalian transcription factor AML1 (named Acute Myeloid Leukemia 1 or Runt-Related Transcription Factor 1) it is suggested that the pathways of cell patterning in the eye resemble those utilized during vertebrate hematopoietic development (Daga et al. [1996]). Very recently, SERPENT was found to cooperate with the cycling transcription factors CLK/CYC in Drosophila "to determine their direct targets and therefore orchestrate tissue-specific clock outputs" (Meireles-Filho et al. [2014]). As the GATA/RUNX interaction during hematopoiesis appears to be conserved in mammals (Waltzer et al. [2003]), a similar mechanism is conceivable there. This is further supported by the macrophage-specific finding of overrepresented binding sites for RUNX factors in circadian genes compared to non-circadian and to non-expressed genes.

Binding sites for transcription factors interacting with GATA proteins are also enriched in promoters of circadian genes. This concerns SP1, the general Ebox binders LMO2, LDB1, TAL-1, and E2A and the ETS family member PU.1 in macrophages as well as the nuclear receptor SF-1, NKX2.5, NFATc4, and MEF2 in liver cells, while all GATA proteins can interact with the histone acetyltransferases P300 and CBP (Molkentin [2000], Ferreira et al. [2005]).

8.7 Conclusion

The thorough choice of background genes for overrepresentation analysis of circadian genes in two mouse cell types led to a deeper understanding of the interplay between transcription factors and promoter sequences in gene regulation. Besides the specific binding sites to which transcription factors bind the surrounding promoter sequence might influence the permanence of the transcription factors' binding presence. Due to altered transcription factor expression levels depending on time and tissue, these features impact the regulators' interactions, contributing to cell type- and phase-specificity of circadian gene expression. The application of overrepresentation analysis to two different background sets differing in their expression properties led to the suggestion of probable transcription factor interactions for regulating ciradian genes in mouse macrophages and liver cells as a starting point for further gene expression modeling.

Appendix A

Motifs listed in the table are bound by transcription factors listed in table 7.1. For each motif the transcription factor class is listed (Stegmaier et al. [2013]) as well as the four FDRs for detecting them in each background comparison mode ($FDR_\alpha(mac, nc)$, $FDR_\alpha(mac, ne)$, $FDR_\alpha(liv, nc)$, $FDR_\alpha(liv, ne)$), the overall $FDR_{\alpha,W}$ for detecting the motif below the significance threshold α=0.03 in any comparison mode and the $FDR_{\alpha,S}$ for detecting the motif in the observed number of comparison modes.

		Mac		**Liv**			
Motif	**TF class**	**nc**	**ne**	**nc**	**ne**	$FDR_{\alpha,W}$	$FDR_{\alpha,S}$
AHRARNT_02	BHLH	0.91	0.91	0.10	0.08	0.01	0.33
AHRHIF_Q6	BHLH	0.02	0.03	0.07	0.36	0.00	0.43
ARNT_01	BHLH	0.19	0.01	0.69	0.08	0.00	0.50
CLOCKBMAL_Q6	BHLH	0.02	0.02	0.14	0.17	0.00	0.31
CMYC_02	BHLH	0.91	0.03	0.67	0.06	0.00	0.44
DEC_Q1	BHLH	0.03	0.91	0.79	0.79	0.02	0.45
E12_Q6	BHLH	0.91	0.91	0.07	0.71	0.04	0.46
E2A_Q6	BHLH	0.19	0.91	0.02	0.70	0.00	0.49
EBOX_Q6_01	BHLH	0.17	0.91	0.19	0.01	0.00	0.40
MAX_01	BHLH	0.91	0.03	0.66	0.04	0.00	0.44
MYCMAX_01	BHLH	0.03	0.02	0.70	0.04	0.00	0.36
MYCMAX_02	BHLH	0.03	0.00	0.65	0.04	0.00	0.40
MYCMAX_03	BHLH	0.91	0.91	0.66	0.09	0.05	0.51
MYC_Q2	BHLH	0.01	0.02	0.24	0.06	0.00	0.32
MYOD_Q6	BHLH	0.02	0.91	0.20	0.76	0.00	0.46
MYOGENIN_Q6	BHLH	0.91	0.91	0.05	0.75	0.03	0.42
NMYC_01	BHLH	0.02	0.01	0.25	0.07	0.00	0.33
PTF1BETA_Q6	BHLH	0.91	0.91	0.66	0.08	0.04	0.50
SREBP1_01	BHLH	0.03	0.91	0.14	0.17	0.00	0.37
STRA13_01	BHLH	0.19	0.01	0.31	0.08	0.00	0.49
USF_01	BHLH	0.91	0.02	0.65	0.03	0.00	0.44
USF2_Q6	BHLH	0.04	0.91	0.25	0.10	0.00	0.41
USF_C	BHLH	0.03	0.91	0.74	0.78	0.02	0.50
USF_Q6	BHLH	0.02	0.01	0.10	0.04	0.00	0.16
CREB_Q3	BZIP	0.91	0.16	0.19	0.16	0.00	0.48
DBP_Q6	BZIP	0.91	0.91	0.17	0.69	0.09	0.52
TAXCREB_02	BZIP	0.19	0.01	0.11	0.15	0.00	0.39
DMRT1_01	DM	0.91	0.03	0.66	0.15	0.00	0.51

		Mac		Liv			
Motif	**TF class**	**nc**	**ne**	**nc**	**ne**	$FDR_{\alpha,W}$	$FDR_{\alpha,S}$
E2F_01	E2F	0.00	0.02	0.73	0.15	0.00	0.39
E2F1DP2_01	E2F	0.02	0.02	0.67	0.74	0.00	0.53
E2F4DP2_01	E2F	0.02	0.02	0.73	0.77	0.00	0.46
FOXO4_01	FORKHEAD	0.91	0.14	0.06	0.01	0.00	0.28
HNF3_Q6	FORKHEAD	0.91	0.91	0.67	0.15	0.08	0.53
WHN_B	FORKHEAD	0.91	0.15	0.19	0.04	0.00	0.41
CP2_01	GRAINY	0.91	0.91	0.66	0.11	0.06	0.52
ALPHACP1_01	HISTONE	0.03	0.02	0.15	0.05	0.00	0.23
CAAT_01	HISTONE	0.03	0.03	0.05	0.04	0.00	0.14
NFY_01	HISTONE	0.91	0.03	0.01	0.01	0.00	0.13
NFY_Q6	HISTONE	0.91	0.91	0.14	0.02	0.00	0.31
NFY_Q6_01	HISTONE	0.19	0.03	0.05	0.01	0.00	0.26
ALX4_01	HOX	0.91	0.91	0.19	0.14	0.02	0.43
CHX10_01	HOX	0.91	0.91	0.70	0.17	0.10	0.52
HNF1_01	HOX	0.91	0.91	0.71	0.17	0.10	0.52
OCT4_01	HOX	0.91	0.91	0.18	0.13	0.02	0.42
OCT4_02	HOX	0.91	0.91	0.16	0.15	0.02	0.41
PAX3_01	HOX	0.19	0.11	0.09	0.75	0.00	0.50
PAX3_B	HOX	0.91	0.91	0.15	0.19	0.02	0.43
PITX2_Q2	HOX	0.91	0.91	0.22	0.18	0.03	0.48
HSF1_Q6	HSF	0.00	0.14	0.66	0.10	0.00	0.49
IRF1_01	IRF	0.91	0.01	0.67	0.15	0.00	0.48
IRF_Q6_01	IRF	0.91	0.04	0.73	0.15	0.00	0.46
MYB_Q5_01	MYB	0.91	0.03	0.76	0.74	0.02	0.51
EBF_Q6	REL	0.91	0.91	0.69	0.15	0.09	0.52
NFKAPPAB50_01	REL	0.91	0.91	0.06	0.76	0.04	0.42
MTATA_B	TBP	0.16	0.91	0.74	0.85	0.09	0.52
BLIMP1_Q6	ZFC2H2	0.15	0.01	0.66	0.15	0.00	0.53
CACBINDPROT_Q6	ZFC2H2	0.91	0.91	0.15	0.31	0.04	0.52
CKROX_Q2	ZFC2H2	0.91	0.13	0.03	0.69	0.00	0.47
EGR2_01	ZFC2H2	0.91	0.19	0.14	0.76	0.02	0.52
GC_01	ZFC2H2	0.91	0.01	0.32	0.04	0.00	0.41
HIC1_03	ZFC2H2	0.91	0.91	0.16	0.71	0.10	0.51
MAZ_Q6	ZFC2H2	0.91	0.91	0.04	0.71	0.02	0.44
NGFIC_01	ZFC2H2	0.91	0.91	0.11	0.77	0.07	0.43
ROAZ_01	ZFC2H2	0.91	0.03	0.08	0.34	0.00	0.46
SP1_Q2_01	ZFC2H2	0.91	0.13	0.16	0.77	0.01	0.48
SP1_Q4_01	ZFC2H2	0.91	0.02	0.15	0.14	0.00	0.35
SP1_Q6_01	ZFC2H2	0.91	0.02	0.65	0.16	0.00	0.52
WT1_Q6	ZFC2H2	0.91	0.91	0.12	0.66	0.07	0.52
ZEC_01	ZFC2H2	0.03	0.03	0.79	0.05	0.00	0.30
BARBIE_01	ZFDOF	0.91	0.91	0.69	0.15	0.08	0.52

Appendix B

Motifs listed in the table are bound by transcription factors listed in table 7.2. For each motif the transcription factor class is listed (Stegmaier et al. [2013]) as well as the four FDRs for detecting them in each background comparison mode ($FDR_\alpha(mac, nc)$, $FDR_\alpha(mac, ne)$, $FDR_\alpha(liv, nc)$, $FDR_\alpha(liv, ne)$), the overall $FDR_{\alpha,W}$ for detecting the motif below the significance threshold α=0.03 in any comparison mode and the $FDR_{\alpha,S}$ for detecting the motif in the observed number of comparison modes.

		Mac		**Liv**			
Motif	**TF class**	**nc**	**ne**	**nc**	**ne**	$FDR_{\alpha,W}$	$FDR_{\alpha,S}$
AHRARNT_01	BHLH	0.91	0.91	0.31	0.76	0.19	0.57
AP4_Q6	BHLH	0.91	0.91	0.31	0.67	0.17	0.62
AP4_Q6_01	BHLH	0.02	0.91	0.31	0.69	0.00	0.58
ARNT_02	BHLH	0.91	0.91	0.66	0.34	0.19	0.64
CMYC_01	BHLH	0.04	0.17	0.66	0.11	0.00	0.53
E2A_Q2	BHLH	0.91	0.91	0.30	0.72	0.18	0.59
E47_01	BHLH	0.02	0.19	0.67	0.72	0.00	0.62
E47_02	BHLH	0.13	0.78	0.66	0.66	0.04	0.71
HAND1E47_01	BHLH	0.91	0.91	0.74	0.34	0.21	0.60
HEN1_02	BHLH	0.91	0.17	0.66	0.66	0.07	0.67
HIF1_Q3	BHLH	0.91	0.91	0.22	0.72	0.13	0.54
HIF1_Q5	BHLH	0.91	0.91	0.67	0.32	0.18	0.62
MYCMAX_B	BHLH	0.91	0.91	0.34	0.77	0.21	0.58
MYOD_01	BHLH	0.91	0.91	0.31	0.77	0.20	0.56
MYOD_Q6_01	BHLH	0.17	0.15	0.77	0.71	0.01	0.62
SREBP1_Q5	BHLH	0.91	0.91	0.34	0.66	0.18	0.64
SREBP1_Q6	BHLH	0.03	0.19	0.67	0.31	0.00	0.63
SREBP_Q3	BHLH	0.19	0.19	0.66	0.70	0.02	0.69
SREBP_Q6	BHLH	0.91	0.91	0.33	0.67	0.18	0.63
TAL1ALPHAE47_01	BHLH	0.91	0.78	0.31	0.31	0.07	0.66
TAL1BETAE47_01	BHLH	0.91	0.03	0.66	0.71	0.01	0.59
TAL1BETAITF2_01	BHLH	0.91	0.03	0.67	0.31	0.01	0.59
TAL1_Q6	BHLH	0.04	0.91	0.31	0.37	0.00	0.62
TFE_Q6	BHLH	0.91	0.19	0.67	0.72	0.08	0.64
USF_02	BHLH	0.91	0.91	0.66	0.18	0.10	0.55
USF_Q6_01	BHLH	0.91	0.91	0.67	0.16	0.09	0.53
AP2ALPHA_02	BHSH	0.78	0.03	0.70	0.75	0.01	0.61
AP2ALPHA_03	BHSH	0.19	0.19	0.66	0.79	0.02	0.66

		Mac		Liv			
Motif	**TF class**	**nc**	**ne**	**nc**	**ne**	$FDR_{\alpha,W}$	$FDR_{\alpha,S}$
AP2_Q6	BHSH	0.91	0.91	0.31	0.71	0.18	0.60
AP1_01	BZIP	0.91	0.17	0.76	0.69	0.08	0.61
AP1_C	BZIP	0.91	0.91	0.66	0.31	0.17	0.63
AP1FJ_Q2	BZIP	0.19	0.91	0.66	0.66	0.07	0.68
AP1_Q2_01	BZIP	0.91	0.91	0.30	0.34	0.08	0.62
AP1_Q4	BZIP	0.91	0.91	0.66	0.31	0.17	0.62
AP1_Q6	BZIP	0.91	0.91	0.73	0.36	0.22	0.61
ATF_01	BZIP	0.91	0.91	0.66	0.32	0.17	0.63
ATF1_Q6	BZIP	0.91	0.19	0.66	0.66	0.07	0.68
ATF3_Q6	BZIP	0.91	0.91	0.66	0.31	0.17	0.63
ATF4_Q2	BZIP	0.13	0.00	0.65	0.67	0.00	0.62
ATF6_01	BZIP	0.91	0.13	0.66	0.67	0.05	0.65
CEBP_01	BZIP	0.91	0.91	0.66	0.18	0.10	0.56
CEBPDELTA_Q6	BZIP	0.17	0.17	0.66	0.66	0.01	0.70
CHOP_01	BZIP	0.91	0.91	0.30	0.71	0.18	0.60
CREB_01	BZIP	0.91	0.91	0.66	0.33	0.18	0.64
CREB_02	BZIP	0.91	0.91	0.30	0.31	0.08	0.61
CREBATF_Q6	BZIP	0.13	0.12	0.23	0.30	0.00	0.59
CREBP1_01	BZIP	0.91	0.04	0.66	0.82	0.02	0.53
CREBP1CJUN_01	BZIP	0.17	0.91	0.66	0.73	0.07	0.64
CREBP1_Q2	BZIP	0.19	0.04	0.73	0.20	0.00	0.55
CREB_Q2	BZIP	0.19	0.02	0.31	0.31	0.00	0.62
CREB_Q2_01	BZIP	0.91	0.91	0.31	0.33	0.08	0.62
CREB_Q4	BZIP	0.78	0.02	0.31	0.33	0.00	0.64
CREB_Q4_01	BZIP	0.91	0.16	0.31	0.31	0.01	0.64
E4BP4_01	BZIP	0.91	0.91	0.67	0.34	0.19	0.64
HLF_01	BZIP	0.91	0.19	0.66	0.66	0.07	0.68
MAF_Q6_01	BZIP	0.91	0.19	0.66	0.66	0.07	0.68
NFE2_01	BZIP	0.19	0.91	0.69	0.72	0.08	0.64
TAXCREB_01	BZIP	0.78	0.02	0.67	0.72	0.01	0.63
VBP_01	BZIP	0.91	0.91	0.66	0.32	0.17	0.63
VJUN_01	BZIP	0.91	0.19	0.18	0.34	0.01	0.60
VMAF_01	BZIP	0.17	0.17	0.31	0.67	0.01	0.68
DMRT4_01	DM	0.91	0.91	0.67	0.37	0.21	0.65
DMRT5_01	DM	0.91	0.91	0.73	0.36	0.21	0.62
DMRT7_01	DM	0.91	0.04	0.66	0.77	0.02	0.56
E2_01	E2	0.91	0.19	0.66	0.34	0.04	0.68
E2_Q6	E2	0.19	0.91	0.67	0.79	0.09	0.61
E2_Q6_01	E2	0.91	0.91	0.66	0.36	0.19	0.65
E2F_02	E2F	0.91	0.91	0.67	0.34	0.19	0.63
E2F_03	E2F	0.78	0.91	0.31	0.24	0.05	0.63
E2F1DP1_01	E2F	0.03	0.17	0.66	0.36	0.00	0.66

		Mac		Liv			
Motif	**TF class**	**nc**	**ne**	**nc**	**ne**	$FDR_{\alpha,W}$	$FDR_{\alpha,S}$
E2F1_Q3	E2F	0.91	0.13	0.66	0.78	0.06	0.60
E2F1_Q3_01	E2F	0.91	0.19	0.71	0.34	0.04	0.65
E2F1_Q4	E2F	0.15	0.78	0.66	0.33	0.03	0.71
E2F1_Q6	E2F	0.91	0.19	0.66	0.77	0.09	0.62
E2F1_Q6_01	E2F	0.91	0.11	0.66	0.72	0.05	0.62
E2F_Q4	E2F	0.02	0.15	0.72	0.34	0.00	0.60
E2F_Q4_01	E2F	0.19	0.91	0.67	0.33	0.04	0.67
E2F_Q6	E2F	0.02	0.19	0.69	0.34	0.00	0.64
E2F_Q6_01	E2F	0.19	0.78	0.73	0.32	0.03	0.68
CETS168_Q6	ETS	0.78	0.01	0.66	0.66	0.01	0.67
CETS1P54_01	ETS	0.78	0.01	0.66	0.66	0.00	0.67
CETS1P54_02	ETS	0.15	0.00	0.66	0.66	0.00	0.63
CETS1P54_03	ETS	0.91	0.03	0.66	0.66	0.01	0.62
ELF1_Q6	ETS	0.91	0.91	0.66	0.37	0.20	0.66
ELK1_02	ETS	0.91	0.02	0.66	0.71	0.01	0.58
ETS_Q4	ETS	0.91	0.19	0.66	0.67	0.07	0.67
ETS_Q6	ETS	0.91	0.02	0.66	0.69	0.01	0.60
GABP_B	ETS	0.19	0.00	0.66	0.67	0.00	0.64
NRF2_01	ETS	0.91	0.02	0.66	0.72	0.01	0.58
PEA3_Q6	ETS	0.91	0.02	0.66	0.66	0.01	0.61
PU1_Q4	ETS	0.91	0.91	0.66	0.15	0.08	0.54
PU1_Q6	ETS	0.91	0.03	0.19	0.63	0.00	0.55
FOXD3_01	FORKHEAD	0.91	0.91	0.69	0.36	0.20	0.63
FOXM1_01	FORKHEAD	0.19	0.19	0.66	0.72	0.02	0.68
FOXO1_01	FORKHEAD	0.91	0.15	0.66	0.08	0.01	0.54
FOXO4_02	FORKHEAD	0.91	0.91	0.66	0.36	0.19	0.65
FREAC2_01	FORKHEAD	0.19	0.19	0.67	0.36	0.01	0.71
FREAC3_01	FORKHEAD	0.91	0.91	0.66	0.33	0.18	0.64
FREAC4_01	FORKHEAD	0.91	0.19	0.67	0.14	0.02	0.57
HFH4_01	FORKHEAD	0.91	0.04	0.73	0.33	0.01	0.57
HNF3_Q6_01	FORKHEAD	0.91	0.91	0.66	0.35	0.19	0.65
CP2_02	GRAINY	0.91	0.91	0.24	0.67	0.13	0.58
LBP1_Q6	GRAINY	0.02	0.91	0.69	0.66	0.01	0.60
ACAAT_B	HISTONE	0.91	0.19	0.72	0.17	0.02	0.56
LEF1_Q2	HMG	0.91	0.91	0.14	0.34	0.04	0.53
LEF1_Q2_01	HMG	0.91	0.91	0.34	0.69	0.19	0.63
SOX9_B1	HMG	0.91	0.91	0.18	0.34	0.05	0.55
SOX_Q6	HMG	0.19	0.19	0.66	0.67	0.01	0.71
SRY_01	HMG	0.91	0.91	0.66	0.31	0.17	0.63
SRY_02	HMG	0.91	0.17	0.66	0.33	0.03	0.66
BRN2_01	HOX	0.91	0.91	0.32	0.16	0.04	0.53
CDPCR1_01	HOX	0.15	0.15	0.66	0.66	0.01	0.69

		Mac		Liv			
Motif	**TF class**	**nc**	**ne**	**nc**	**ne**	$FDR_{\alpha,W}$	$FDR_{\alpha,S}$
CDPCR3HD_01	HOX	0.13	0.17	0.31	0.83	0.01	0.59
CRX_Q4	HOX	0.15	0.19	0.30	0.71	0.01	0.66
HMX1_01	HOX	0.03	0.91	0.34	0.72	0.01	0.58
HNF1_Q6_01	HOX	0.91	0.91	0.65	0.16	0.08	0.55
HOXA3_01	HOX	0.91	0.91	0.66	0.34	0.18	0.64
HOXA7_01	HOX	0.91	0.91	0.31	0.19	0.05	0.54
IPF1_Q4	HOX	0.91	0.91	0.19	0.66	0.10	0.56
MEF3_B	HOX	0.91	0.91	0.31	0.22	0.06	0.55
NANOG_01	HOX	0.91	0.91	0.66	0.35	0.19	0.65
NKX25_01	HOX	0.78	0.14	0.32	0.69	0.02	0.69
NKX25_Q5	HOX	0.78	0.91	0.30	0.66	0.14	0.68
OCT1_Q5_01	HOX	0.91	0.91	0.67	0.15	0.08	0.53
PAX5_01	HOX	0.91	0.91	0.31	0.77	0.20	0.56
PAX8_B	HOX	0.19	0.91	0.66	0.66	0.07	0.68
PAX9_B	HOX	0.91	0.91	0.74	0.38	0.23	0.62
PBX1_02	HOX	0.91	0.19	0.69	0.77	0.09	0.61
TGIF_01	HOX	0.91	0.03	0.66	0.72	0.01	0.58
TTF1_Q6	HOX	0.78	0.91	0.30	0.35	0.08	0.68
XVENT1_01	HOX	0.19	0.19	0.71	0.14	0.00	0.59
HSF_Q6	HSF	0.03	0.91	0.67	0.67	0.01	0.60
ICSBP_Q6	IRF	0.91	0.91	0.66	0.15	0.08	0.54
IRF2_01	IRF	0.91	0.17	0.67	0.11	0.01	0.55
IRF7_01	IRF	0.17	0.91	0.64	0.67	0.07	0.67
IRF_Q6	IRF	0.91	0.91	0.66	0.15	0.08	0.54
ISRE_01	IRF	0.17	0.03	0.66	0.14	0.00	0.54
MEF2_Q6_01	MADS	0.12	0.91	0.73	0.74	0.06	0.57
RSRFC4_01	MADS	0.19	0.91	0.66	0.66	0.07	0.68
SRF_C	MADS	0.91	0.91	0.31	0.79	0.20	0.55
SRF_Q4	MADS	0.91	0.91	0.22	0.70	0.13	0.55
SRF_Q5_01	MADS	0.91	0.91	0.31	0.72	0.18	0.59
SRF_Q5_02	MADS	0.91	0.19	0.66	0.66	0.07	0.68
SRF_Q6	MADS	0.91	0.91	0.31	0.76	0.20	0.57
VMYB_02	MYB	0.91	0.02	0.66	0.74	0.01	0.57
BEL1_B	NA	0.91	0.91	0.30	0.67	0.17	0.62
GCM_Q2	NA	0.91	0.19	0.72	0.66	0.08	0.65
LDSPOLYA_B	NA	0.19	0.91	0.32	0.31	0.02	0.65
NRF1_Q6	NA	0.78	0.13	0.72	0.71	0.05	0.65
OLF1_01	NA	0.17	0.16	0.66	0.80	0.01	0.63
POLY_C	NA	0.91	0.04	0.71	0.72	0.02	0.55
TFIII_Q6	NA	0.91	0.91	0.15	0.66	0.08	0.54
P53_01	P53	0.91	0.91	0.67	0.33	0.18	0.63
P53_02	P53	0.91	0.15	0.71	0.30	0.03	0.62

		Mac		Liv			
Motif	**TF class**	**nc**	**ne**	**nc**	**ne**	$FDR_{\alpha,W}$	$FDR_{\alpha,S}$
CREL_01	REL	0.15	0.17	0.71	0.83	0.01	0.59
NFAT_Q4_01	REL	0.91	0.91	0.66	0.15	0.08	0.54
NFAT_Q6	REL	0.91	0.91	0.66	0.15	0.08	0.54
NFKAPPAB_01	REL	0.13	0.02	0.16	0.65	0.00	0.54
NFKAPPAB65_01	REL	0.12	0.10	0.67	0.81	0.01	0.58
NFKB_C	REL	0.91	0.19	0.71	0.17	0.02	0.56
NFKB_Q6	REL	0.15	0.19	0.67	0.66	0.01	0.70
NFKB_Q6_01	REL	0.12	0.13	0.67	0.85	0.01	0.57
RBPJK_Q4	REL	0.00	0.00	0.66	0.66	0.00	0.57
EFC_Q6	RFX	0.91	0.91	0.67	0.32	0.18	0.62
AML1_01	RUNT	0.91	0.19	0.66	0.66	0.07	0.68
AML_Q6	RUNT	0.91	0.17	0.66	0.72	0.07	0.64
OSF2_Q6	RUNT	0.19	0.17	0.66	0.66	0.01	0.70
PEBP_Q6	RUNT	0.91	0.17	0.66	0.72	0.07	0.64
NF1_Q6	SMAD	0.03	0.17	0.66	0.18	0.00	0.56
SMAD3_Q6	SMAD	0.11	0.91	0.66	0.32	0.02	0.64
SMAD_Q6_01	SMAD	0.13	0.17	0.66	0.73	0.01	0.65
STAT_01	STAT	0.91	0.91	0.66	0.14	0.07	0.53
STAT1_01	STAT	0.91	0.91	0.66	0.16	0.09	0.55
STAT3_01	STAT	0.91	0.13	0.66	0.24	0.02	0.60
STAT3_02	STAT	0.91	0.15	0.66	0.76	0.07	0.61
TATA_01	TBP	0.19	0.91	0.32	0.78	0.04	0.61
TATA_C	TBP	0.19	0.91	0.67	0.67	0.08	0.67
TBX5_Q5	TBX	0.91	0.19	0.73	0.73	0.09	0.60
TEF1_Q6	TEA	0.91	0.78	0.67	0.14	0.07	0.59
E4F1_Q6	ZFC2H2	0.16	0.02	0.33	0.15	0.00	0.54
EGR3_01	ZFC2H2	0.91	0.91	0.31	0.66	0.17	0.62
EGR_Q6	ZFC2H2	0.91	0.91	0.15	0.66	0.08	0.54
GFI1_Q6	ZFC2H2	0.91	0.15	0.67	0.77	0.07	0.60
GZF1_01	ZFC2H2	0.78	0.14	0.69	0.75	0.06	0.66
IK1_01	ZFC2H2	0.15	0.91	0.66	0.66	0.06	0.67
IK3_01	ZFC2H2	0.91	0.91	0.66	0.33	0.18	0.63
IK_Q5	ZFC2H2	0.91	0.17	0.66	0.73	0.07	0.63
MOVOB_01	ZFC2H2	0.91	0.91	0.14	0.66	0.07	0.53
MTF1_Q4	ZFC2H2	0.19	0.17	0.65	0.71	0.01	0.69
NFMUE1_Q6	ZFC2H2	0.91	0.12	0.67	0.77	0.05	0.59
RP58_01	ZFC2H2	0.12	0.14	0.67	0.67	0.01	0.66
SP1_Q6	ZFC2H2	0.91	0.02	0.30	0.72	0.00	0.55
SP3_Q3	ZFC2H2	0.91	0.12	0.67	0.67	0.05	0.64
SPZ1_01	ZFC2H2	0.91	0.15	0.65	0.72	0.07	0.64
STAF_01	ZFC2H2	0.91	0.17	0.66	0.66	0.07	0.67
STAF_02	ZFC2H2	0.19	0.91	0.66	0.67	0.07	0.67

		Mac		Liv			
Motif	**TF class**	**nc**	**ne**	**nc**	**ne**	$FDR_{\alpha,W}$	$FDR_{\alpha,S}$
SZF11_01	ZFC2H2	0.91	0.91	0.34	0.66	0.19	0.64
YY1_02	ZFC2H2	0.19	0.91	0.66	0.66	0.07	0.68
YY1_Q6	ZFC2H2	0.91	0.03	0.66	0.79	0.01	0.54
YY1_Q6_02	ZFC2H2	0.91	0.91	0.66	0.15	0.08	0.54
ZBRK1_01	ZFC2H2	0.03	0.13	0.31	0.73	0.00	0.57
ZID_01	ZFC2H2	0.11	0.13	0.71	0.66	0.01	0.64
AR_01	ZFC4-NR	0.19	0.04	0.76	0.34	0.00	0.61
AR_02	ZFC4-NR	0.91	0.04	0.67	0.37	0.01	0.63
AR_03	ZFC4-NR	0.91	0.91	0.67	0.35	0.19	0.64
ARP1_01	ZFC4-NR	0.15	0.91	0.66	0.66	0.06	0.67
AR_Q2	ZFC4-NR	0.19	0.91	0.66	0.82	0.09	0.60
AR_Q6	ZFC4-NR	0.91	0.17	0.65	0.78	0.08	0.62
COUP_01	ZFC4-NR	0.91	0.91	0.67	0.31	0.17	0.62
DR3_Q4	ZFC4-NR	0.91	0.03	0.66	0.67	0.01	0.61
DR4_Q2	ZFC4-NR	0.19	0.19	0.66	0.77	0.02	0.66
FXR_Q3	ZFC4-NR	0.91	0.91	0.67	0.31	0.17	0.62
GRE_C	ZFC4-NR	0.19	0.04	0.66	0.67	0.00	0.65
GR_Q6	ZFC4-NR	0.91	0.91	0.65	0.19	0.10	0.57
GR_Q6_01	ZFC4-NR	0.91	0.91	0.66	0.22	0.12	0.58
HNF4ALPHA_Q6	ZFC4-NR	0.91	0.91	0.66	0.31	0.17	0.62
HNF4_Q6_01	ZFC4-NR	0.91	0.91	0.66	0.18	0.10	0.56
LXR_DR4_Q3	ZFC4-NR	0.91	0.04	0.66	0.65	0.01	0.63
PPARA_02	ZFC4-NR	0.91	0.91	0.32	0.69	0.18	0.62
PPAR_DR1_Q2	ZFC4-NR	0.91	0.91	0.66	0.33	0.18	0.64
PPARG_01	ZFC4-NR	0.19	0.91	0.71	0.82	0.10	0.57
PR_02	ZFC4-NR	0.91	0.91	0.66	0.16	0.09	0.54
PXRRXR_02	ZFC4-NR	0.91	0.19	0.66	0.72	0.08	0.65
RORA2_01	ZFC4-NR	0.91	0.91	0.66	0.16	0.09	0.54
SF1_Q6_01	ZFC4-NR	0.91	0.91	0.18	0.69	0.10	0.53
T3R_01	ZFC4-NR	0.91	0.91	0.66	0.34	0.19	0.64
GATA1_01	ZFGATA	0.19	0.91	0.31	0.73	0.04	0.63
GATA1_03	ZFGATA	0.14	0.15	0.30	0.30	0.00	0.65
GATA1_04	ZFGATA	0.91	0.19	0.66	0.31	0.04	0.67
GATA1_05	ZFGATA	0.91	0.03	0.69	0.69	0.01	0.58
GATA2_01	ZFGATA	0.91	0.91	0.31	0.75	0.19	0.58
GATA3_01	ZFGATA	0.17	0.91	0.66	0.80	0.08	0.60

Bibliography

U. Abraham, A. E. Granada, P. O. Westermark, M. Heine, A. Kramer, and H. Herzel. Coupling governs entrainment range of circadian clocks. *Mol Syst Biol*, 6:438, Nov 2010. doi: 10.1038/msb.2010.92. URL http://dx.doi.org/10.1038/msb.2010.92.

S. Adhikary and M. Eilers. Transcriptional regulation and transformation by MYC proteins. *Nat Rev Mol Cell Biol*, 6(8):635–645, Aug 2005. doi: 10.1038/nrm1703. URL http://dx.doi.org/10.1038/nrm1703.

L. Aguilar-Arnal and P. Sassone-Corsi. The circadian epigenome: how metabolism talks to chromatin remodeling. *Curr Opin Cell Biol*, 25(2):170–176, Apr 2013. doi: 10.1016/j.ceb.2013.01.003. URL http://dx.doi.org/10.1016/j.ceb.2013.01.003.

H. Akbulut, F. Icli, A. Büyükcelik, K. G. Akbulut, and S. Demirci. The role of granulocyte-macrophage-colony stimulating factor, cortisol, and melatonin in the regulation of the circadian rhythms of peripheral blood cells in healthy volunteers and patients with breast cancer. *J Pineal Res*, 26(1):1–8, Jan 1999. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1600-079X.1999.tb00560.x/pdf.

U. Alon. Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8(6):450–461, Jun 2007. doi: 10.1038/nrg2102. URL http://dx.doi.org/10.1038/nrg2102.

G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723, Jun 2009. doi: 10.1126/science.1162327. URL http://dx.doi.org/10.1126/science.1162327.

A. Balsalobre. Clock genes in mammalian peripheral tissues. *Cell Tissue Res*, 309(1):193–199, Jul 2002. doi: 10.1007/s00441-002-0585-0. URL http://dx.doi.org/10.1007/s00441-002-0585-0.

A. R. Barnard and P. M. Nolan. When clocks go bad: neurobehavioural consequences of disrupted circadian timing. *PLoS Genet*, 4(5):e1000040, May 2008. doi: 10.1371/journal.pgen.1000040. URL http://dx.doi.org/10.1371/journal.pgen.1000040.

M. M. Bellet and P. Sassone-Corsi. Mammalian circadian clock and metabolism - the epigenetic link. *J Cell Sci*, 123(Pt 22):3837–3848, Nov 2010. doi: 10.1242/jcs.051649. URL http://dx.doi.org/10.1242/jcs.051649.

O. Bembom. *seqLogo: Sequence logos for DNA sequence alignments*, 1990. URL `http://www.bioconductor.org/packages/release/bioc/vignettes/seqLogo/inst/doc/seqLogo.pdf`. R package version 1.20.0.

Y. Benjamini and M. Bogomolov. Adjusting for selection bias in testing multiple families of hypotheses. *The Annals of Applied Statistics*, June 2011. doi: arXiv:1106.3670. URL `http://arxiv.org/abs/1106.3670v1`.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series*, 57:289–300, 1995. URL `http://www.jstor.org/stable/2346101`.

O. Bonny, M. Vinciguerra, M. L. Gumz, and G. Mazzoccoli. Molecular bases of circadian rhythmicity in renal physiology and pathology. *Nephrol Dial Transplant*, 28(10):2421–2431, Oct 2013. doi: 10.1093/ndt/gft319. URL `http://dx.doi.org/10.1093/ndt/gft319`.

K. Bozek, S. M. Kiełbasa, A. Kramer, and H. Herzel. Promoter analysis of mammalian clock controlled genes. *Genome Inform*, 18:65–74, 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/18546475`.

K. Bozek, A. Relógio, S. M. Kiełbasa, M. Heine, C. Dame, A. Kramer, and H. Herzel. Regulation of clock-controlled genes in mammals. *PLoS One*, 4(3):e4882, 2009. doi: 10.1371/journal.pone.0004882. URL `http://dx.doi.org/10.1371/journal.pone.0004882`.

K. Bozek, A. L. Rosahl, S. Gaub, S. Lorenzen, and H. Herzel. Circadian transcription in liver. *Biosystems*, 102(1):61–69, Oct 2010. doi: 10.1016/j.biosystems.2010.07.010. URL `http://dx.doi.org/10.1016/j.biosystems.2010.07.010`.

S. A. Brown, E. Kowalska, and R. Dallmann. (Re)inventing the circadian feedback loop. *Dev Cell*, 22(3):477–487, Mar 2012. doi: 10.1016/j.devcel.2012.02.007. URL `http://dx.doi.org/10.1016/j.devcel.2012.02.007`.

E. D. Buhr and J. S. Takahashi. *Molecular components of the Mammalian circadian clock*. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-25950-0_1. URL `http://dx.doi.org/10.1007/978-3-642-25950-0_1`.

A. Burger, A. M. Walczak, and P. G. Wolynes. Abduction and asylum in the lives of transcription factors. *Proc Natl Acad Sci U S A*, 107(9):4016–4021, Mar 2010. doi: 10.1073/pnas.0915138107. URL `http://dx.doi.org/10.1073/pnas.0915138107`.

B. R. Cairns. The logic of chromatin architecture and remodelling at promoters. *Nature*, 461(7261):193–198, Sep 2009. doi: 10.1038/nature08450. URL `http://dx.doi.org/10.1038/nature08450`.

O. Castanon-Cervantes, M. Wu, J. C. Ehlen, K. Paul, K. L. Gamble, R. L. Johnson, R. C. Besing, M. Menaker, A. T. Gewirtz, and A. J. Davidson. Dysregulation of

inflammatory responses by chronic circadian disruption. *J Immunol*, 185(10):5796–5805, Nov 2010. doi: 10.4049/jimmunol.1001026. URL http://dx.doi.org/10.4049/jimmunol.1001026.

L.-W. Chang, R. Nagarajan, J. A. Magee, J. Milbrandt, and G. D. Stormo. A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. *Genome Res*, 16(3):405–413, Mar 2006. doi: 10.1101/gr.4303406. URL http://dx.doi.org/10.1101/gr.4303406.

E. N. C. O. D. E. P. Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, Jun 2007. doi: 10.1038/nature05874. URL http://dx.doi.org/10.1038/nature05874.

t. W. I. o. S. Crown Human Genome Center, Department of Molecular Genetics. Genecards, 2013. URL http://www.genecards.org/.

A. Daga, C. A. Karlovich, K. Dumstrei, and U. Banerjee. Patterning of cells in the drosophila eye by Lozenge, which shares homologous domains with AML1. *Genes Dev*, 10(10):1194–1205, May 1996. URL http://www.ncbi.nlm.nih.gov/pubmed/8675007.

C. Dibner, U. Schibler, and U. Albrecht. The mammalian circadian timing system: organization and coordination of central and peripheral clocks. *Annu Rev Physiol*, 72:517–549, 2010. doi: 10.1146/annurev-physiol-021909-135821. URL http://dx.doi.org/10.1146/annurev-physiol-021909-135821.

C. J. Doherty and S. A. Kay. Circadian control of global gene expression patterns. *Annu Rev Genet*, 44:419–444, 2010. doi: 10.1146/annurev-genet-102209-163432. URL http://dx.doi.org/10.1146/annurev-genet-102209-163432.

M. Doi, J. Hirayama, and P. Sassone-Corsi. Circadian regulator clock is a histone acetyltransferase. *Cell*, 125(3):497–508, May 2006. doi: 10.1016/j.cell.2006.03.033. URL http://dx.doi.org/10.1016/j.cell.2006.03.033.

J. C. Dunlap. Molecular bases for circadian clocks. *Cell*, 96(2):271–290, Jan 1999. URL http://www.ncbi.nlm.nih.gov/pubmed/9988221.

D. J. Durgan and M. E. Young. The cardiomyocyte circadian clock: emerging roles in health and disease. *Circ Res*, 106(4):647–658, Mar 2010. doi: 10.1161/CIRCRESAHA.109.209957. URL http://dx.doi.org/10.1161/CIRCRESAHA.109.209957.

R. S. Edgar, E. W. Green, Y. Zhao, G. van Ooijen, M. Olmedo, X. Qin, Y. Xu, M. Pan, U. K. Valekunja, K. A. Feeney, E. S. Maywood, M. H. Hastings, N. S. Baliga, M. Merrow, A. J. Millar, C. H. Johnson, C. P. Kyriacou, J. S. O'Neill, and A. B. Reddy. Peroxiredoxins are conserved markers of circadian rhythms. *Nature*, 485(7399):459–464, May 2012. doi: 10.1038/nature11088. URL http://dx.doi.org/10.1038/nature11088.

B. Efron. Simultaneous inference: when should hypothesis testing problems be combined? *The Annals of Applied Statistics*, 2:197–223, 2008. doi: 10.1214/07-AOAS141. URL `http://arxiv.org/abs/0803.3863`.

M. Eilers and R. N. Eisenman. Myc's broad reach. *Genes Dev*, 22(20):2755–2766, Oct 2008. doi: 10.1101/gad.1712408. URL `http://dx.doi.org/10.1101/gad.1712408`.

M. A. Elhelu. The role of macrophages in immunology. *J Natl Med Assoc*, 75(3):314–317, Mar 1983. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2561478/`.

R. Ferreira, K. Ohneda, M. Yamamoto, and S. Philipsen. GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol Cell Biol*, 25(4):1215–1227, Feb 2005. doi: 10.1128/MCB.25.4.1215-1227.2005. URL `http://dx.doi.org/10.1128/MCB.25.4.1215-1227.2005`.

R. Ferreira, A. Wai, R. Shimizu, N. Gillemans, R. Rottier, M. von Lindern, K. Ohneda, F. Grosveld, M. Yamamoto, and S. Philipsen. Dynamic regulation of Gata factor levels is more important than their identity. *Blood*, 109(12):5481–5490, Jun 2007. doi: 10.1182/blood-2006-11-060491. URL `http://dx.doi.org/10.1182/blood-2006-11-060491`.

J. D. Fleming, G. Pavesi, P. Benatti, C. Imbriano, R. Mantovani, and K. Struhl. NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. *Genome Res*, 23(8):1195–1209, Aug 2013. doi: 10.1101/gr.148080.112. URL `http://dx.doi.org/10.1101/gr.148080.112`.

P. Fonjallaz, V. Ossipow, G. Wanner, and U. Schibler. The two PAR leucine zipper proteins, TEF and DBP, display similar circadian and tissue-specific expression, but have different target promoter preferences. *EMBO J*, 15(2):351–362, Jan 1996. URL `http://www.ncbi.nlm.nih.gov/pubmed/8617210`.

J. J. Gagliardino, M. T. Pessacq, R. E. Hernandez, and O. R. Rebolledo. Circadian variations in serum glucagon and hepatic glycogen and cyclic amp concentrations. *J Endocrinol*, 78(2):297–298, Aug 1978. URL `http://www.ncbi.nlm.nih.gov/pubmed/212498`.

J. E. Gale, H. I. Cox, J. Qian, G. D. Block, C. S. Colwell, and A. V. Matveyenko. Disruption of circadian rhythms accelerates development of diabetes through pancreatic beta-cell loss and dysfunction. *J Biol Rhythms*, 26(5):423–433, Oct 2011. doi: 10.1177/0748730411416341. URL `http://dx.doi.org/10.1177/0748730411416341`.

D. Gatfield and U. Schibler. Circadian glucose homeostasis requires compensatory interference between brain and liver clocks. *Proc Natl Acad Sci U S A*, 105(39):14753–14754, Sep 2008. doi: 10.1073/pnas.0807861105. URL `http://dx.doi.org/10.1073/pnas.0807861105`.

V. Gobert, D. Osman, S. Bras, B. Augé, M. Boube, H.-M. Bourbon, T. Horn, M. Boutros, M. Haenlin, and L. Waltzer. A genome-wide RNA interference screen identifies a differential role of the mediator CDK8 module subunits for GATA/ RUNX-activated transcription in drosophila. *Mol Cell Biol*, 30(11):2837–2848, Jun 2010. doi: 10.1128/MCB.01625-09. URL `http://dx.doi.org/10.1128/MCB.01625-09`.

R. Gordân, N. Shen, I. Dror, T. Zhou, J. Horton, R. Rohs, and M. L. Bulyk. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep*, 3(4):1093–1104, Apr 2013. doi: 10.1016/j.celrep.2013.03.014. URL `http://dx.doi.org/10.1016/j.celrep.2013.03.014`.

V. Guasconi, H. Yahi, and S. Ait-Si-Ali. Transcription factors. *Atlas of Genetics and Cytogenetics in Oncology and Haematology*, 2003. URL `http://AtlasGeneticsOncology.org/Educ/TFactorsEng.html`.

G. L. Hager, J. G. McNally, and T. Misteli. Transcription dynamics. *Mol Cell*, 35(6):741–753, Sep 2009. doi: 10.1016/j.molcel.2009.09.005. URL `http://dx.doi.org/10.1016/j.molcel.2009.09.005`.

F. Halberg, E. A. Johnson, B. W. Brown, and J. J. Bittner. Susceptibility rhythm to E. coli endotoxin and bioassay. *Proceedings of The Society for Experimental Biology and Medicine*, 103:142, Feb 1960. doi: 10.3181/00379727-103-25439. URL `http://www.ncbi.nlm.nih.gov/pubmed/14398944`.

P. E. Hardin and W. Yu. Circadian transcription: passing the HAT to CLOCK. *Cell*, 125(3):424–426, May 2006. doi: 10.1016/j.cell.2006.04.010. URL `http://dx.doi.org/10.1016/j.cell.2006.04.010`.

B. Herkert and M. Eilers. Transcriptional repression: the dark side of myc. *Genes Cancer*, 1(6):580–586, Jun 2010. doi: 10.1177/1947601910379012. URL `http://dx.doi.org/10.1177/1947601910379012`.

H. Hosoda, K. Kato, H. Asano, M. Ito, H. Kato, T. Iwamoto, A. Suzuki, S. Masushige, and S. Kida. CBP/p300 is a cell type-specific modulator of CLOCK/BMAL1-mediated transcription. *Mol Brain*, 2:34, 2009. doi: 10.1186/1756-6606-2-34. URL `http://dx.doi.org/10.1186/1756-6606-2-34`.

Y. F. Hu, B. Lüscher, A. Admon, N. Mermod, and R. Tjian. Transcription factor AP-4 contains multiple dimerization domains that regulate dimer specificity. *Genes Dev*, 4(10):1741–1752, Oct 1990. URL `http://www.ncbi.nlm.nih.gov/pubmed/2123466`.

M. E. Hughes, L. DiTacchio, K. R. Hayes, C. Vollmers, S. Pulivarthy, J. E. Baggs, S. Panda, and J. B. Hogenesch. Harmonics of circadian gene transcription in mammals. *PLoS Genet*, 5(4):e1000442, Apr 2009. doi: 10.1371/journal.pgen.1000442. URL `http://dx.doi.org/10.1371/journal.pgen.1000442`.

D. A. Hume. The biology of macrophages - an online review. *online*, 2012. URL `http://www.macrophages.com/macrophage-review`.

R. A. Hut and L. Steyvers. Circwavebatch version 3.3, January 2007. URL `http://www.rug.nl/fwn/onderzoek/programmas/biologie/chronobiologie/downloads/index`.

S. Jacobshagen, B. Kessler, and C. A. Rinehart. At least four distinct circadian regulatory mechanisms are required for all phases of rhythms in mRNA amount. *J Biol Rhythms*, 23(6):511–524, Dec 2008. doi: 10.1177/0748730408325753. URL `http://dx.doi.org/10.1177/0748730408325753`.

S. Jayanthi and D. D. Vecchio. Tuning genetic clocks employing DNA binding sites. *PLoS One*, 7(7):e41019, 2012. doi: 10.1371/journal.pone.0041019. URL `http://dx.doi.org/10.1371/journal.pone.0041019`.

S. Jayanthi, K. S. Nilgiriwala, and D. D. Vecchio. Retroactivity controls the temporal dynamics of gene transcription. *ACS Synth Biol*, 2(8):431–441, Aug 2013. doi: 10.1021/sb300098w. URL `http://dx.doi.org/10.1021/sb300098w`.

C. H. Johnson. Circadian clocks and cell division: what's the pacemaker? *Cell Cycle*, 9(19):3864–3873, Oct 2010. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3047750/`.

S. Jones. An overview of the basic helix-loop-helix proteins. *Genome Biol*, 5(6):226, 2004. doi: 10.1186/gb-2004-5-6-226. URL `http://dx.doi.org/10.1186/gb-2004-5-6-226`.

T. Kadesch. Consequences of heteromeric interactions among helix-loop-helix proteins. *Cell Growth Differ*, 4(1):49–55, Jan 1993. URL `http://www.ncbi.nlm.nih.gov/pubmed/8424906`.

M. T. Kassouf, J. R. Hughes, S. Taylor, S. J. McGowan, S. Soneji, A. L. Green, P. Vyas, and C. Porcher. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res*, 20(8):1064–1083, Aug 2010. doi: 10.1101/gr.104935.110. URL `http://dx.doi.org/10.1101/gr.104935.110`.

A. Kel, O. Kel-Margoulis, A. Romaschenko, E. Wingender, and V. Ratner. Composite Modules - The DNA Blueprints of Combinatorial Transcriptional Regulation in Multicellular Organisms. In *BGRS2000. Proceedings of the Second International Conference on Bioinformatics of Genome Regulation and Structure*, pages 123–126, 2000. URL `http://www.researchgate.net/publication/259645651`.

M. Keller, J. Mazuch, U. Abraham, G. D. Eom, E. D. Herzog, H.-D. Volk, A. Kramer, and B. Maier. A circadian clock in macrophages controls inflammatory immune responses. *Proc Natl Acad Sci U S A*, 106(50):21407–21412, Dec 2009. doi: 10.1073/pnas.0906361106. URL `http://dx.doi.org/10.1073/pnas.0906361106`.

S. M. Kiełbasa, H. Klein, H. G. Roider, M. Vingron, and N. Blüthgen. Transfind–predicting transcriptional regulators for gene sets. *Nucleic Acids Res*, 38(Web Server issue):W275–W280, Jul 2010. doi: 10.1093/nar/gkq438. URL `http://dx.doi.org/10.1093/nar/gkq438`.

N. Koike, S.-H. Yoo, H.-C. Huang, V. Kumar, C. Lee, T.-K. Kim, and J. S. Takahashi. Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science*, 338(6105):349–354, Oct 2012. doi: 10.1126/science.1226339. URL `http://dx.doi.org/10.1126/science.1226339`.

A. Korenčič, G. Bordyugov, R. Košir, D. Rozman, M. Goličnik, and H. Herzel. The interplay of cis-regulatory elements rules circadian rhythms in mouse liver. *PLoS One*, 7(11):e46835, 2012. doi: 10.1371/journal.pone.0046835. URL `http://dx.doi.org/10.1371/journal.pone.0046835`.

B. Kornmann, O. Schaad, H. Bujard, J. S. Takahashi, and U. Schibler. System-driven and oscillator-dependent circadian transcription in mice with a conditionally active liver clock. *PLoS Biol*, 5(2):e34, Feb 2007a. doi: 10.1371/journal.pbio.0050034. URL `http://dx.doi.org/10.1371/journal.pbio.0050034`.

B. Kornmann, O. Schaad, H. Reinke, C. Saini, and U. Schibler. Regulation of circadian gene expression in liver by systemic signals and hepatocyte oscillators. *Cold Spring Harb Symp Quant Biol*, 72:319–330, 2007b. doi: 10.1101/sqb.2007.72.041. URL `http://dx.doi.org/10.1101/sqb.2007.72.041`.

G. Krucik, 2013. URL `http://www.healthline.com/human-body-maps/liver`.

J. M. Landolin, D. S. Johnson, N. D. Trinklein, S. F. Aldred, C. Medina, H. Shulha, Z. Weng, and R. M. Myers. Sequence features that drive human promoter function and tissue specificity. *Genome Res*, 20(7):890–898, Jul 2010. doi: 10.1101/gr.100370.109. URL `http://dx.doi.org/10.1101/gr.100370.109`.

E. Laurenti, B. Varnum-Finney, A. Wilson, I. Ferrero, W. E. Blanco-Bose, A. Ehninger, P. S. Knoepfler, P.-F. Cheng, H. R. MacDonald, R. N. Eisenman, I. D. Bernstein, and A. Trumpp. Hematopoietic stem cell function and survival depend on c-Myc and N-Myc activity. *Cell Stem Cell*, 3(6):611–624, Dec 2008. doi: 10.1016/j.stem.2008.09.005. URL `http://dx.doi.org/10.1016/j.stem.2008.09.005`.

D. Levens. You Don't Muck with MYC. *Genes Cancer*, 1(6):547–554, Jun 2010. doi: 10.1177/1947601910377492. URL `http://dx.doi.org/10.1177/1947601910377492`.

A. J. Lewy, R. L. Sack, M. L. Blood, V. K. Bauer, N. L. Cutler, and K. H. Thomas. Melatonin marks circadian phase position and resets the endogenous circadian pacemaker in humans. *Ciba Found Symp*, 183:303–17; discussion 317–21, 1995. URL `http://www.ncbi.nlm.nih.gov/pubmed/7656692`.

S. Li, B. Aufiero, R. L. Schiltz, and M. J. Walsh. Regulation of the homeodomain CCAAT displacement/cut protein function by histone acetyltransferases p300/CREB-binding protein (CBP)-associated factor and CBP. *Proc Natl Acad Sci U S A*, 97(13):7166–7171, Jun 2000. doi: 10.1073/pnas.130028697. URL `http://dx.doi.org/10.1073/pnas.130028697`.

D. Liu, S. D. Peddada, L. Li, and C. R. Weinberg. Phase analysis of circadian-related genes in two tissues. *BMC Bioinformatics*, 7:87, 2006a. doi: 10.1186/1471-2105-7-87. URL `http://dx.doi.org/10.1186/1471-2105-7-87`.

J. Liu, G. Malkani, G. Mankani, X. Shi, M. Meyer, S. Cunningham-Runddles, X. Ma, and Z. S. Sun. The circadian clock Period 2 gene regulates gamma interferon production of NK cells in host response to lipopolysaccharide-induced endotoxic shock. *Infect Immun*, 74(8):4750–4756, Aug 2006b. doi: 10.1128/IAI.00287-06. URL `http://dx.doi.org/10.1128/IAI.00287-06`.

T. Manke, H. G. Roider, and M. Vingron. Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS Comput Biol*, 4(3):e1000039, Mar 2008. doi: 10.1371/journal.pcbi.1000039. URL `http://dx.doi.org/10.1371/journal.pcbi.1000039`.

G. L. Martelot, D. Canella, L. Symul, E. Migliavacca, F. Gilardi, R. Liechti, O. Martin, K. Harshman, M. Delorenzi, B. Desvergne, W. Herr, B. Deplancke, U. Schibler, J. Rougemont, N. Guex, N. Hernandez, F. Naef, and C. Consortium. Genome-wide RNA polymerase II profiles and RNA accumulation reveal kinetics of transcription and associated epigenetic changes during diurnal cycles. *PLoS Biol*, 10(11): e1001442, 2012. doi: 10.1371/journal.pbio.1001442. URL `http://dx.doi.org/10.1371/journal.pbio.1001442`.

S. Masri and P. Sassone-Corsi. Plasticity and specificity of the circadian epigenome. *Nat Neurosci*, 13(11):1324–1329, Nov 2010. doi: 10.1038/nn.2668. URL `http://dx.doi.org/10.1038/nn.2668`.

M. E. Massari and C. Murre. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol Cell Biol*, 20(2):429–440, Jan 2000. URL `http://www.ncbi.nlm.nih.gov/pubmed/10611221`.

V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue): D108–D110, Jan 2006. doi: 10.1093/nar/gkj143. URL `http://dx.doi.org/10.1093/nar/gkj143`.

E. S. Maywood, J. E. Chesham, J. A. O'Brien, and M. H. Hastings. A diversity of paracrine signals sustains molecular circadian cycling in suprachiasmatic nu-

cleus circuits. *Proc Natl Acad Sci U S A*, 108(34):14306–14311, Aug 2011. doi: 10.1073/pnas.1101767108. URL http://dx.doi.org/10.1073/pnas.1101767108.

A. C. A. Meireles-Filho, A. F. Bardet, J. O. Yánez-Cuna, G. Stampfel, and A. Stark. cis-Regulatory Requirements for Tissue-Specific Programs of the Circadian Clock. *Curr Biol*, 24(1):1–10, Jan 2014. doi: 10.1016/j.cub.2013.11.017. URL http://dx.doi.org/10.1016/j.cub.2013.11.017.

J. S. Menet, S. Pescatore, and M. Rosbash. CLOCK:BMAL1 is a pioneer-like transcription factor. *Genes Dev*, 28(1):8–13, Jan 2014. doi: 10.1101/gad.228536.113. URL http://dx.doi.org/10.1101/gad.228536.113.

G. Meng, A. Mosig, and M. Vingron. A computational evaluation of over-representation of regulatory motifs in the promoter regions of differentially expressed genes. *BMC Bioinformatics*, 11:267, 2010. doi: 10.1186/1471-2105-11-267. URL http://dx.doi.org/10.1186/1471-2105-11-267.

J. A. Mohawk, C. B. Green, and J. S. Takahashi. Central and peripheral circadian clocks in mammals. *Annu Rev Neurosci*, 35:445–462, 2012. doi: 10.1146/annurev-neuro-060909-153128. URL http://dx.doi.org/10.1146/annurev-neuro-060909-153128.

J. D. Molkentin. The zinc finger-containing transcription factors GATA-4, -5, and -6. Ubiquitously expressed regulators of tissue-specific gene expression. *J Biol Chem*, 275 (50):38949–38952, Dec 2000. doi: 10.1074/jbc.R000029200. URL http://dx.doi.org/10.1074/jbc.R000029200.

E. Munoz and R. Baler. The circadian E-box: when perfect is not good enough. *Chronobiol Int*, 20(3):371–388, May 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/12868535.

E. Munoz, M. Brewer, and R. Baler. Modulation of BMAL/CLOCK/E-Box complex activity by a CT-rich cis-acting element. *Mol Cell Endocrinol*, 252(1-2):74–81, Jun 2006. doi: 10.1016/j.mce.2006.03.007. URL http://dx.doi.org/10.1016/j.mce.2006.03.007.

Y. Nakahata, M. Kaluzova, B. Grimaldi, S. Sahar, J. Hirayama, D. Chen, L. P. Guarente, and P. Sassone-Corsi. The NAD+-dependent deacetylase SIRT1 modulates CLOCK-mediated chromatin remodeling and circadian control. *Cell*, 134(2):329–340, Jul 2008. doi: 10.1016/j.cell.2008.07.002. URL http://dx.doi.org/10.1016/j.cell.2008.07.002.

Y. Nakahata, S. Sahar, G. Astarita, M. Kaluzova, and P. Sassone-Corsi. Circadian control of the NAD+ salvage pathway by CLOCK-SIRT1. *Science*, 324(5927):654–657, May 2009. doi: 10.1126/science.1170803. URL http://dx.doi.org/10.1126/science.1170803.

G. Natoli. Little things that count in transcriptional regulation. *Cell*, 118(4):406–408, Aug 2004. doi: 10.1016/j.cell.2004.08.003. URL `http://dx.doi.org/10.1016/j.cell.2004.08.003`.

M. Nichols, F. Weih, W. Schmid, C. DeVack, E. Kowenz-Leutz, B. Luckow, M. Boshart, and G. Schütz. Phosphorylation of CREB affects its binding to high and low affinity sites: implications for cAMP induced gene transcription. *EMBO J*, 11(9):3337–3346, Sep 1992. URL `http://www.ncbi.nlm.nih.gov/pubmed/1354612`.

Z. Nie, G. Hu, G. Wei, K. Cui, A. Yamane, W. Resch, R. Wang, D. R. Green, L. Tessarollo, R. Casellas, K. Zhao, and D. Levens. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, 151(1):68–79, Sep 2012. doi: 10.1016/j.cell.2012.08.033. URL `http://dx.doi.org/10.1016/j.cell.2012.08.033`.

K. Nowick and L. Stubbs. Lineage-specific transcription factors and the evolution of gene regulatory networks. *Brief Funct Genomics*, 9(1):65–78, Jan 2010. doi: 10.1093/bfgp/elp056. URL `http://dx.doi.org/10.1093/bfgp/elp056`.

K.-J. Oh, H.-S. Han, M.-J. Kim, and S.-H. Koo. CREB and FoxO1: two transcription factors for the regulation of hepatic gluconeogenesis. *BMB Rep*, 46(12):567–574, Dec 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/24238363`.

M. Pachkov, I. Erb, N. Molina, and E. van Nimwegen. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res*, 35(Database issue): D127–D131, Jan 2007. doi: 10.1093/nar/gkl857. URL `http://dx.doi.org/10.1093/nar/gkl857`.

R. Padinhateeri and J. F. Marko. Nucleosome positioning in a model of active chromatin remodeling enzymes. *Proc Natl Acad Sci U S A*, 108(19):7799–7803, May 2011. doi: 10.1073/pnas.1015206108. URL `http://dx.doi.org/10.1073/pnas.1015206108`.

C. L. Partch and K. H. Gardner. Coactivator recruitment: a new role for PAS domains in transcriptional regulation by the bHLH-PAS family. *J Cell Physiol*, 223(3):553–557, Jun 2010. doi: 10.1002/jcp.22067. URL `http://dx.doi.org/10.1002/jcp.22067`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

S. Rahmann, T. Müller, and M. Vingron. On the power of profiles for transcription factor binding site detection. *Stat Appl Genet Mol Biol*, 2:Article7, 2003. doi: 10.2202/1544-6115.1032. URL `http://dx.doi.org/10.2202/1544-6115.1032`.

M. R. Ralph, R. G. Foster, F. C. Davis, and M. Menaker. Transplanted suprachiasmatic nucleus determines circadian period. *Science*, 247(4945):975–978, Feb 1990. URL `http://www.ncbi.nlm.nih.gov/pubmed/2305266`.

T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, P. Carninci, C. O. Daub, A. R. R. Forrest, J. Gough, S. Grimmond, J.-H. Han, T. Hashimoto, W. Hide, O. Hofmann, A. Kamburov, M. Kaur, H. Kawaji, A. Kubosaki, T. Lassmann, E. van Nimwegen, C. R. MacPherson, C. Ogawa, A. Radovanovic, A. Schwartz, R. D. Teasdale, J. Tegnér, B. Lenhard, S. A. Teichmann, T. Arakawa, N. Ninomiya, K. Murakami, M. Tagami, S. Fukuda, K. Imamura, C. Kai, R. Ishihara, Y. Kitazume, J. Kawai, D. A. Hume, T. Ideker, and Y. Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, Mar 2010. doi: 10.1016/j.cell.2010.01.044. URL `http://dx.doi.org/10.1016/j.cell.2010.01.044`.

M. Raymondjean, S. Cereghini, and M. Yaniv. Several distinct 'CCAAT' box binding proteins coexist in eukaryotic cells. *Proc Natl Acad Sci U S A*, 85(3):757–761, Feb 1988. URL `http://www.ncbi.nlm.nih.gov/pubmed/3422457`.

J. S. Rest, K. Bullaughey, G. P. Morris, and W.-H. Li. Contribution of transcription factor binding site motif variants to condition-specific gene expression patterns in budding yeast. *PLoS One*, 7(2):e32274, 2012. doi: 10.1371/journal.pone.0032274. URL `http://dx.doi.org/10.1371/journal.pone.0032274`.

G. Rey, F. Cesbron, J. Rougemont, H. Reinke, M. Brunner, and F. Naef. Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. *PLoS Biol*, 9(2):e1000595, Feb 2011. doi: 10.1371/journal.pbio.1000595. URL `http://dx.doi.org/10.1371/journal.pbio.1000595`.

J. Richards and M. L. Gumz. Advances in understanding the peripheral circadian clocks. *FASEB J*, 26(9):3602–3613, Sep 2012. doi: 10.1096/fj.12-203554. URL `http://dx.doi.org/10.1096/fj.12-203554`.

J. A. Ripperger and M. Merrow. Perfect timing: epigenetic regulation of the circadian clock. *FEBS Lett*, 585(10):1406–1411, May 2011. doi: 10.1016/j.febslet.2011.04.047. URL `http://dx.doi.org/10.1016/j.febslet.2011.04.047`.

H. G. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, Jan 2007. doi: 10.1093/bioinformatics/btl565. URL `http://dx.doi.org/10.1093/bioinformatics/btl565`.

H. G. Roider, B. Lenhard, A. Kanhere, S. A. Haas, and M. Vingron. CpG-depleted promoters harbor tissue-specific transcription factor binding signals–implications for motif overrepresentation analyses. *Nucleic Acids Res*, 37(19):6305–6315, Oct 2009a. doi: 10.1093/nar/gkp682. URL `http://dx.doi.org/10.1093/nar/gkp682`.

H. G. Roider, T. Manke, S. O'Keeffe, M. Vingron, and S. A. Haas. PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, 25(4):435–442, Feb 2009b. doi: 10.1093/bioinformatics/btn627. URL `http://dx.doi.org/10.1093/bioinformatics/btn627`.

J. Rosenblatt. A practitioner's guide to multiple testing error rates. *The Annals of Applied Statistics*, Jun 2013. doi: arXiv:1304.4920. URL `http://arxiv-web3.library.cornell.edu/abs/1304.4920`.

S. Sahar and P. Sassone-Corsi. Circadian rhythms and memory formation: regulation by chromatin remodeling. *Front Mol Neurosci*, 5:37, 2012. doi: 10.3389/fnmol.2012.00037. URL `http://dx.doi.org/10.3389/fnmol.2012.00037`.

A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–D94, Jan 2004. doi: 10.1093/nar/gkh012. URL `http://dx.doi.org/10.1093/nar/gkh012`.

S. Sandoval, M. Pigazzi, and K. M. Sakamoto. CREB: A Key Regulator of Normal and Neoplastic Hematopoiesis. *Adv Hematol*, 2009:634292, 2009. doi: 10.1155/2009/634292. URL `http://dx.doi.org/10.1155/2009/634292`.

K. Sankaran. Multiple Testing for Hierarchically Dependent Hypotheses: Assessment and Application of Recent Methods, 2011. URL `http://web.stanford.edu/~kriss1/uploads/1/3/7/7/13777035/technical_report_2011_kris_vigre_8_26_2011.pdf`.

C. Savvidis and M. Koutsilieris. Circadian rhythm disruption in cancer biology. *Mol Med*, 18:1249–1260, 2012. doi: 10.2119/molmed.2012.00077. URL `http://dx.doi.org/10.2119/molmed.2012.00077`.

M. Schaub, A. Krol, and P. Carbon. Structural organization of Staf-DNA complexes. *Nucleic Acids Res*, 28(10):2114–2121, May 2000. URL `http://www.ncbi.nlm.nih.gov/pubmed/10773080`.

K. Schellenberg. Molecular mechanism of circadian erythropoietin regulation. Master's thesis, Charité Master program Molecular Medicine, 2008.

T. Schlake, M. Schorpp, M. Nehls, and T. Boehm. The nude gene encodes a sequence-specific DNA binding protein with homologs in organisms that lack an anticipatory immune system. *Proc Natl Acad Sci U S A*, 94(8):3842–3847, Apr 1997. URL `http://www.ncbi.nlm.nih.gov/pubmed/9108066`.

E. Segal and J. Widom. What controls nucleosome positions? *Trends Genet*, 25(8):335–343, Aug 2009. doi: 10.1016/j.tig.2009.06.002. URL `http://dx.doi.org/10.1016/j.tig.2009.06.002`.

S. SethuNarayanan. Role of the DBP Gene in the Regulation of Circadian and Cyclic Hematopoiesis: A Case for Potential Linkages. *Int J Hum Genet*, 11(3):135–147, 2011. URL `http://www.krepublishers.com/02-Journals/IJHG/IJHG-11-0-000-11-Web/IJHG-11-3-000-11-Abst-PDF/IJHG-11-3-135-11-460-Narayanan-S-R-S/IJHG-11-3-135-11-460-Narayanan-S-R-S-Tt.pdf`.

C. E. Shannon. The mathematical theory of communication. 1963. *MD Comput*, 14(4):306–317, 1997. URL `http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf`.

X. Shi, C. C. Metges, and H.-M. Seyfert. Interaction of C/EBP-beta and NF-Y factors constrains activity levels of the nutritionally controlled promoter IA expressing the acetyl-CoA carboxylase-alpha gene in cattle. *BMC Mol Biol*, 13:21, 2012. doi: 10.1186/1471-2199-13-21. URL `http://dx.doi.org/10.1186/1471-2199-13-21`.

K. Shimomura, V. Kumar, N. Koike, T.-K. Kim, J. Chong, E. D. Buhr, A. R. Whiteley, S. S. Low, C. Omura, D. Fenner, J. R. Owens, M. Richards, S.-H. Yoo, H.-K. Hong, M. H. Vitaterna, J. Bass, M. T. Pletcher, T. Wiltshire, J. Hogenesch, P. L. Lowrey, and J. S. Takahashi. Usf1, a suppressor of the circadian Clock mutant, reveals the nature of the DNA-binding of the CLOCK:BMAL1 complex in mice. *Elife*, 2:e00426, 2013. doi: 10.7554/eLife.00426. URL `http://dx.doi.org/10.7554/eLife.00426`.

M. L. Spengler, K. K. Kuropatwinski, M. Comas, A. V. Gasparian, N. Fedtsova, A. S. Gleiberman, I. I. Gitlin, N. M. Artemicheva, K. A. Deluca, A. V. Gudkov, and M. P. Antoch. Core circadian protein CLOCK is a positive regulator of NF-κB-mediated transcription. *Proc Natl Acad Sci U S A*, 109(37):E2457–E2465, Sep 2012. doi: 10.1073/pnas.1206274109. URL `http://dx.doi.org/10.1073/pnas.1206274109`.

P. Stegmaier, A. E. Kel, and E. Wingender. Systematic DNA-binding domain classification of transcription factors. *Genome Inform*, 15(2):276–286, 2004. URL `http://www.bioinf.med.uni-goettingen.de/fileadmin/upload/publications/GIW04_44_Stegmaier.pdf`.

P. Stegmaier, A. Kel, E. Wingender, and J. Borlak. A discriminative approach for unsupervised clustering of DNA sequence motifs. *PLoS Comput Biol*, 9(3):e1002958, 2013. doi: 10.1371/journal.pcbi.1002958. URL `http://dx.doi.org/10.1371/journal.pcbi.1002958`.

K.-F. Storch, O. Lipan, I. Leykin, N. Viswanathan, F. C. Davis, W. H. Wong, and C. J. Weitz. Extensive and divergent circadian gene expression in liver and heart. *Nature*, 417(6884):78–83, May 2002. doi: 10.1038/nature744. URL `http://dx.doi.org/10.1038/nature744`.

H. Taniguchi, A. F. Fernández, F. Setién, S. Ropero, E. Ballestar, A. Villanueva, H. Yamamoto, K. Imai, Y. Shinomura, and M. Esteller. Epigenetic inactivation of the circadian clock gene BMAL1 in hematologic malignancies. *Cancer Res*, 69(21):8447–8454, Nov 2009. doi: 10.1158/0008-5472.CAN-09-0551. URL `http://dx.doi.org/10.1158/0008-5472.CAN-09-0551`.

M. Thomas-Chollier, A. Hufton, M. Heinig, S. O'Keeffe, N. E. Masri, H. G. Roider, T. Manke, and M. Vingron. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat Protoc*, 6(12):1860–1869,

Dec 2011. doi: 10.1038/nprot.2011.409. URL http://dx.doi.org/10.1038/nprot.2011.409.

H. R. Ueda, S. Hayashi, W. Chen, M. Sano, M. Machida, Y. Shigeyoshi, M. Iino, and S. Hashimoto. System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nat Genet*, 37(2):187–192, Feb 2005. doi: 10.1038/ng1504. URL http://dx.doi.org/10.1038/ng1504.

T. Ueshima, T. Kawamoto, K. K. Honda, M. Noshiro, K. Fujimoto, S. Nakao, N. Ichinose, S. Hashimoto, O. Gotoh, and Y. Kato. Identification of a new clock-related element EL-box involved in circadian regulation by BMAL1/CLOCK and HES1. *Gene*, 510(2):118–125, Dec 2012. doi: 10.1016/j.gene.2012.08.022. URL http://dx.doi.org/10.1016/j.gene.2012.08.022.

M. Ukai-Tadenuma, R. G. Yamada, H. Xu, J. A. Ripperger, A. C. Liu, and H. R. Ueda. Delay in feedback repression by cryptochrome 1 is required for circadian clock function. *Cell*, 144(2):268–281, Jan 2011. doi: 10.1016/j.cell.2010.12.019. URL http://dx.doi.org/10.1016/j.cell.2010.12.019.

E. Valen and A. Sandelin. Genomic and chromatin signals underlying transcription start-site selection. *Trends Genet*, 27(11):475–485, Nov 2011. doi: 10.1016/j.tig.2011.08.001. URL http://dx.doi.org/10.1016/j.tig.2011.08.001.

A. Valera, A. Pujol, X. Gregori, E. Riu, J. Visa, and F. Bosch. Evidence from transgenic mice that myc regulates hepatic glycolysis. *FASEB J*, 9(11):1067–1078, Aug 1995. URL http://www.fasebj.org/content/9/11/1067.

K. Vanselow, J. T. Vanselow, P. O. Westermark, S. Reischl, B. Maier, T. Korte, A. Herrmann, H. Herzel, A. Schlosser, and A. Kramer. Differential effects of PER2 phosphorylation: molecular basis for the human familial advanced sleep phase syndrome (FASPS). *Genes Dev*, 20(19):2660–2672, Oct 2006. doi: 10.1101/gad.397006. URL http://dx.doi.org/10.1101/gad.397006.

J. Vervoorts, J. Lüscher-Firzlaff, and B. Lüscher. The ins and outs of MYC regulation by posttranslational mechanisms. *J Biol Chem*, 281(46):34725–34729, Nov 2006. doi: 10.1074/jbc.R600017200. URL http://dx.doi.org/10.1074/jbc.R600017200.

M. Vignali, A. H. Hassan, K. E. Neely, and J. L. Workman. ATP-dependent chromatin-remodeling complexes. *Mol Cell Biol*, 20(6):1899–1910, Mar 2000. URL http://mcb.asm.org/content/20/6/1899.full.

A. E. Vinogradov. Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. *Trends Genet*, 21(12):639–643, Dec 2005. doi: 10.1016/j.tig.2005.09.002. URL http://dx.doi.org/10.1016/j.tig.2005.09.002.

T. Wallach, K. Schellenberg, B. Maier, R. K. R. Kalathur, P. Porras, E. E. Wanker, M. E. Futschik, and A. Kramer. Dynamic circadian protein-protein interaction networks predict temporal organization of cellular functions. *PLoS Genet*, 9(3):e1003398, Mar 2013. doi: 10.1371/journal.pgen.1003398. URL `http://dx.doi.org/10.1371/journal.pgen.1003398`.

L. Waltzer, G. Ferjoux, L. Bataillé, and M. Haenlin. Cooperation between the GATA and RUNX factors Serpent and Lozenge during Drosophila hematopoiesis. *EMBO J*, 22(24):6516–6525, Dec 2003. doi: 10.1093/emboj/cdg622. URL `http://dx.doi.org/10.1093/emboj/cdg622`.

N. Wang, G. Yang, Z. Jia, H. Zhang, T. Aoyagi, S. Soodvilai, J. D. Symons, J. B. Schnermann, F. J. Gonzalez, S. E. Litwin, and T. Yang. Vascular PPARgamma controls circadian variation in blood pressure and heart rate through Bmal1. *Cell Metab*, 8(6):482–491, Dec 2008. doi: 10.1016/j.cmet.2008.10.009. URL `http://dx.doi.org/10.1016/j.cmet.2008.10.009`.

P. O. Westermark and H. Herzel. Mechanism for 12 hr rhythm generation by the circadian clock. *Cell Rep*, 3(4):1228–1238, Apr 2013. doi: 10.1016/j.celrep.2013.03.013. URL `http://dx.doi.org/10.1016/j.celrep.2013.03.013`.

I. Wierstra and J. Alves. The c-myc promoter: still MysterY and challenge. *Adv Cancer Res*, 99:113–333, 2008. doi: 10.1016/S0065-230X(07)99004-1. URL `http://dx.doi.org/10.1016/S0065-230X(07)99004-1`.

J. Xiao, Y. Zhou, H. Lai, S. Lei, L. H. Chi, and X. Mo. Transcription Factor NF-Y Is a Functional Regulator of the Transcription of Core Clock Gene Bmal1. *J Biol Chem*, 288(44):31930–31936, Nov 2013. doi: 10.1074/jbc.M113.507038. URL `http://dx.doi.org/10.1074/jbc.M113.507038`.

Y. Xu, Y. L. Zhou, W. Luo, Q.-S. Zhu, D. Levy, O. A. MacDougald, and M. L. Snead. NF-Y and CCAAT/enhancer-binding protein alpha synergistically activate the mouse amelogenin gene. *J Biol Chem*, 281(23):16090–16098, Jun 2006. doi: 10.1074/jbc.M510514200. URL `http://dx.doi.org/10.1074/jbc.M510514200`.

T. Yamamoto, Y. Nakahata, H. Soma, M. Akashi, T. Mamine, and T. Takumi. Transcriptional oscillation of canonical clock genes in mouse peripheral tissues. *BMC Mol Biol*, 5:18, Oct 2004. doi: 10.1186/1471-2199-5-18. URL `http://dx.doi.org/10.1186/1471-2199-5-18`.

J. Yan, H. Wang, Y. Liu, and C. Shao. Analysis of gene regulatory networks in the mammalian circadian rhythm. *PLoS Comput Biol*, 4(10):e1000193, Oct 2008. doi: 10.1371/journal.pcbi.1000193. URL `http://dx.doi.org/10.1371/journal.pcbi.1000193`.

R. Yang and Z. Su. Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics*, 26(12):i168–i174, Jun 2010. doi: 10.1093/bioinformatics/btq189. URL `http://dx.doi.org/10.1093/bioinformatics/btq189`.

D. Yekutieli. Hierarchical False Discovery Rate-Controlling Methodology. *Journal of the American Statistical Association*, 103:309–316, 2008. URL `http://dx.doi.org/10.1198/016214507000001373`.

D. Yekutieli and Y. Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82:171–196, Dec 1999. URL `http://www.sciencedirect.com/science/article/pii/S0378375899000415`.

D. Yekutieli, A. Reiner-Benaim, Y. Benjamini, G. I. Elmer, N. Kafkafi, N. E. Letwin, and N. H. Lee. Approaches to multiplicity issues in complex research in microarray analysis. *Statistica Neerlandica*, 60(4):414–437, 2006. URL `http://EconPapers.repec.org/RePEc:bla:stanee:v:60:y:2006:i:4:p:414-437`.

L. Yin and M. A. Lazar. The orphan nuclear receptor Rev-erbalpha recruits the N-CoR/histone deacetylase 3 corepressor to regulate the circadian Bmal1 gene. *Mol Endocrinol*, 19(6):1452–1459, Jun 2005. doi: 10.1210/me.2005-0057. URL `http://dx.doi.org/10.1210/me.2005-0057`.

K. S. Zaret and J. S. Carroll. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev*, 25(21):2227–2241, Nov 2011. doi: 10.1101/gad.176826.111. URL `http://dx.doi.org/10.1101/gad.176826.111`.

X. Zhang, D. T. Odom, S.-H. Koo, M. D. Conkright, G. Canettieri, J. Best, H. Chen, R. Jenner, E. Herbolsheimer, E. Jacobsen, S. Kadam, J. R. Ecker, B. Emerson, J. B. Hogenesch, T. Unterman, R. A. Young, and M. Montminy. Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. *Proc Natl Acad Sci U S A*, 102(12):4459–4464, Mar 2005. doi: 10.1073/pnas.0501076102. URL `http://dx.doi.org/10.1073/pnas.0501076102`.

R. Zheng, B. Rebolledo-Jaramillo, Y. Zong, L. Wang, P. Russo, W. Hancock, B. Z. Stanger, R. C. Hardison, and G. A. Blobel. Function of GATA Factors in the Adult Mouse Liver. *PLoS One*, 8(12):e83723, 2013. doi: 10.1371/journal.pone.0083723. URL `http://dx.doi.org/10.1371/journal.pone.0083723`.

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe. Darüberhinaus trugen klärende Gespräche mit Arbeitsgruppenmitgliedern und auf Konferenzen zur Ergebnisfindung bei. Ich habe mich weder an einem anderen Ort um einen Doktorgrad beworben, noch besitze ich bereits einen entsprechenden Titel. Die dem angestrebten Verfahren zugrunde liegende Promotionsordnung ist mir bekannt.

Berlin, den 25.3.2014 Agnes Lioba Rosahl

List of publications

K. Bozek, A. L. Rosahl, S. Gaub, S. Lorenzen, and H. Herzel. Circadian transcription in liver. Biosystems, 102(1):61-69, Oct 2010. doi: 10.1016/j.biosystems.2010.07.010.